The 2 x 2 x 2 case in causality, of an effect, a cause and a confounder. A cross-over guide to the 2 x 2 x 2 contingency table

Thomas Colignatus

Thomas Cool Consultancy & Econometrics

30. May 2007

The 2 × 2 × 2 contingency table in causality with an effect, a cause and a confounder

A cross-over guide

Thomas Colignatus, June 19 2007

http://www.dataweb.nl/~cool

Summary

Basic causality is that a cause is present or absent and that the effect follows with a success or not. This happy state of affairs becomes opaque when there is a third variable that can be present or absent and that might be a seeming cause. The 2 × 2 × 2 layout deserves the standard name of the ETC contingency table, with variables Effect, Truth and Confounding and values {S, -S}, {C, -C}, {F, -F}. Assuming the truth we can find the impact of the cause from when the confounder is absent. The 8 cells in the crosstable can be fully parameterized and the conditions for a proper cause can be formulated, with the parameters interpretable as regression coefficients. Requiring conditional independence would be too strong since it neglects some causal processes. The Simpson paradox will not occur if logical consistency is required rather than conditional independence. The paper gives a taxonomy of issues of confounding, a parameterization by risk or safety, and develops the various cases of dependence and (conditional) independence. The paper is supported by software that allows variations. The paper has been written by an econometrician used to structural equations models but visiting epidemiology hoping to use those techniques in experimental economics.

Table of contents

Introduction

Experimental economics may sometimes borrow techniques from epidemiology but the different fields may use different terminologies so that the translation may not be too easy. There may also be hidden assumptions that make sense in one field but no sense in the other one. The following is a case in point. Schield (1999, 2003), "Simpson's paradox and Cornfield's conditions", gives an interesting if not illuminating discussion on causality versus confounding that would be of general interest to economics as well. However, after closer investigation there appear to be some assumptions that probably were all too obvious for the epidemiologists but that at first escaped the attention of this author who is a mere econometrician trained in "structural equations modelling". The general setting is interesting in itself. Pearl (2000) explains that economics has a long tradition of handling causality and indeed using those structural equations models. Thus both fields of study handle causality, as all scientists must. Yet, the different conventions and uses of language can still cause problems of translation. The following tries to bridge the communication gap. The following discussion thus is only for scientists who cross over. Scientists working in only their own field of study and not the other may not be particularly enticed by this effort at translation since they will not experience any problems in communication. Also, this article has been written by an economist and the examples will be from epidemiology. The discussion is directed at the fellow cross-overs or those potentially interested. Economists who have never seen the terminology before may need to exercise some patience. And epidemiologists crossing over might be abhorred by this economic look at their subject.

The simplest case in causality and confounding is when the variables for effect, causality and confounding have only 2 values each, i.e. "present" versus "absent", which gives a 2 × 2 × 2 contingency table. The data are mere counts. Thus we have nominal data collected in a contingency table. Since it is useful to have easy mnenomics and since "causality" and "confounding" both start with a "c", the causality variable will be called "truth". Hence the standard layout is the ETC crosstable with *Effect, Truth* and *Confounding* as the *variables,* and with entries {Success, ¬Success}, {Cause, ¬Cause} and {Confounder, ¬Confounder} as the *values* that the variables can take.

For clarity and completeness it must be emphasized that this discussion thus excludes the 2 × 2 × 2 contingency table where there would be one cause and two common effects, or one effect and two confounders, or two effects and one confounder, or just three variables whatsoever where the researcher is merely interested in some association. Also excluded is the 2 × 2 × 2 contingency table where the effect is an incidence count

and another dimension contains person-years so that those two dimensions are merely used to calculate the proper effect variable of incidence rates: since that table comes down to the 2 × 2 case.

It is also assumed that the causal relations are rather simple. The question on the table (in a double sense) is just the direct line from the other two variables to the effect, all other possibilities excluded. Normally the researcher has a theory of the problem at hand and this should help the researcher to determine the direction of causation. For example, when the Central Bank raises the rate of interest then the mortgage rate will usually rise as well but it is less likely that when you personally switch to another mortgager with a lower rate that the Central Bank will follow too. In the relation between smoking and lung cancer, it may be that the effect (lung cancer) may cause people to smoke, but theory will suggest that this is not the most likely order of events except for a few cases where it indeed might happen in that way. It may also be that all three variables derive from a joint common cause. But that would introduce a fourth variable and that is not the current problem.

The problem setting is that the researcher has no easy way to determine the time sequence of events. The data may be aggregated over time so that all sequential detail might be lost. The effect is unquestionable but there would be doubt about the cause. We may apply the table to a randomized controlled trial but we might also apply it to an observational study. The limitations to intervention can be practical or moral, in that you would not willingly subject an economy to huge inflation or unemployment, or subject a person to some disease. Hence cause and confounder are observed simultaneously and the key question is whether the statistical proportions allow us to determine which of the two is the *true* cause. With $Y = E$ as the variable that must be explained and the explanatory variables X and Z, then one tries both ETC = EXZ and ETC = EZX, and sees whether the statistical proportions give the confounder away. In this paper we tackle this problem by assuming that we know the true cause and then see whether it indeed can be detected. Our approach is logical, in that we analyze the data as they are, in terms of categories and properties, and we don't consider the question that the quantities or properties are so close together that we would need assumptions on theoretical statistical distributions. We call the confounder the "confounder" since we have to look to the situation where it is absent to find the true impact of the cause.

In itself, the exposition below might win in clarity if we first discuss the case where there is no confounder and then later add the case where the confounder exists (i.e. can be present or not). This may very well be the format that is eventually chosen. For the

discussion below we however follow the format that all three variables exist so that it will take some effort to go from the average results to the true causes below it.

We will take the position of a student who is used to the 2 × 2 table and who is suddenly exposed to the shock of a third variable. Our position is a bit like the reader of Kleinbaum et al. (2003), "ActivEpi". Studying this book, the student has been using 2 × 2 tables for 9 chapters and then in chapter 10 suddenly meets a third variable. The shock might cause that the student doesn't understand anything anymore. Have we been studying averages or the controlled subcase where the confounder is absent ? What is the causal model ? Why is the cause called the confounder ? The student has been creating explanations and assumptions of himself or herself for 9 chapters to make sense of the analyses but suddenly is confronted with confounding, which is not only the title of chapter 10 but also the apt description of the student's new state of mind. For epidemiologists these 2 × 2 × 2 tables may be rather complex and not quickly discussed, or, those might be seen as too simple, only drawn at home but not quickly stated in papers for fear of appearing simplistic. Admittedly, issues of didactics and clarity are probably personal to a high degree. It would seem though that clarity might increase in general if the 2 × 2 × 2 table already was discussed in chapter 2 and was more often stated in the journals. The following discussion is rather long and thus is not a suggestion how such a chapter 2 might look like for books for cross-over economists, but it will contain suggestions to that effect.

Appendix A discusses the Simpson paradox with reference to Schield (2003), "Simpson' s paradox and Cornfield's conditions", and **Appendix B** discusses Schield's example of Sir R.A. Fisher on smoking and confounding, and Cornfield's conditions. The conditions by Cornfield et al. are sufficient to block a Simpson paradox but they may be too strong. **Appendix C** derives the parameters of the crosstable using safety parameters. **Appendix D** derives the parameters of the table using average risk parameters. **Appendix E** discusses the risk difference and the Schield plot. **Appendix F** discusses an example from Pearl (2000) that uses two causes instead of one. **Appendix G** uses our results to review chapter 10 in the Kleinbaum et al. (2003). **Appendix H** contains a note on how a chapter 2 might look. **Appendix I** lists the routines in *Mathematica*.

This paper is part of the project Colignatus (2007e), "Elementary statistics and causality" (ESAC). The present paper is a report on how the author has come to understand issues. ESAC itself will be written from a didactic point of view, where the road to understanding is irrelevant.

The 2 × 2 × 2 case

The basic causal model is a 2 × 2 × 2 contingency table in the order *Effect*, *Truth*, *Confounding* (ETC), where *Truth* reflects the true cause and *Confounding* a true confounder. We assume to know the truth and nothing but the truth so we need not worry about whether things are different than stated. The following is a purely theoretical numerical example.

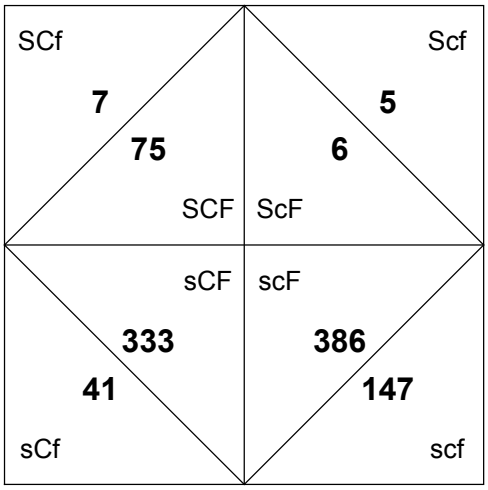
CT[Default, "ETC"]

		Cause	¬ Cause
Success	Confounder	75	6
	¬ Confounder	7	5
¬ Success	Confounder	333	386
	¬ Confounder	41	147

Note the nomenclature. We already mentioned the point in the Introduction and repeat it here because it appears to be rather important. For a real variable x we are used that it has values like 3.45 and 1006.4, and we may easily write P(y | x) for the conditional probability and be sloppy about the distinction between the random variable x and the values that it takes along the real axis. Now for nominal variable *Effect* we have {Success, ¬Success}, for *Truth* we have {Cause, ¬Cause} and for *Confounding* we have values {Confounder, ¬ Confounder}. We can take just single letter symbols too. For identification we take the letter "F" for "ConFounder". Thus we have variables *E, T, C(-ing)* versus values *S, C, F* and 8 combinations with their negations. But now it makes quite a difference whether we write P[E | T] or P[S | C] since the first concerns the variables while the second concerns just two of the values that can be taken.

The numbers and their labels can be shown in a square, where we use capitals for presence and lower case letters for absence. The rows and columns are like in the table but the confounder takes the inner diamond. It is conceivable to adjust the sizes of the surfaces for the actual weights but this has not been programmed and we just print the numbers.

```
ETCSquare[];
```



A major result of this paper is that we can decompose above contingency table into the total number of observations n and 7 driving parameters. We can show this now so that the remainder of the paper will be needed to explain those parameters.

```
lis = SafetyToETCArray[{c, r, b}, {f, q}, {w, v}];
```

```
TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	$(c-(1-f)\,q)\,(1-w)$	$(-c+f\,(1-q)+q)\,(1-v)$
	¬ Confounder	$(1-f)\,q\,r$	$b\,(1-f)\,(1-q)$
¬ Success	Confounder	$(c-(1-f)\,q)\,w$	$(-c+f\,(1-q)+q)\,v$
	¬ Confounder	$(1-f)\,q\,(1-r)$	$(1-b)\,(1-f)\,(1-q)$

The probabilities of success, conditional on truth or confounding, are (transposing to keep the cause in the columns), and noting that b is the background risk:

```
TableForm[lis[[1]] / (lis[[1]] + lis[[2]]) // Simplify // Transpose,
  TableHeadings → {{Confounder, ¬ Confounder}, {Cause, ¬ Cause}}]
```

	Cause	¬ Cause
Confounder	$1-w$	$1-v$
¬ Confounder	r	b

A taxonomy of confounding

There appears to be a rather sizeable literature on trying to find a good definition for what "true confounding" is, see for example Pearl (1998), that has become Chapter 6 in Pearl (2000). The alternative approach is not to worry about "true confounding" but to create a list of all possible sources for confusion. We actually need a taxonomy of confounding like we already have books on logical fallacies, with catching names like "post hoc ergo propter hoc" (she took a medicine and cured), "ad hominem" (playing the person and not the ball), "petitio principii" (begging the question), etcetera. Above we mentioned already the core reason why we called one variable confounding, i.e. that we by assumption *know* that one variable with one value is the true cause so that the other variable must be confounding, so that to find the impact of the cause we have to consider the situation where the confounder is absent. If we didn't have this certainty then we might be unsure about the order EXZ or EZX ("confusus directionis"). But we might also be unsure about the size of the effect ("confusus magnitudinis"). We can also identify some other confusions as well: (a) mixing up the ETC analysis with other kinds of problems in 2 × 2 × 2 tables ("confusus definitionis"), such as the case of one cause and two effects, (b) mixing up variables and values ("confusus nomenclaturis"). Below we will meet some other aspects where we can be confounded on. We will collect them in a basket and at the end of the paper we will turn over the basket and count our treasure.

One example of the "confusus definitionis" might arise for us when we consider Kleinbaum et al. (2003), the chapter 10 on confounding. In their approach, the F and not-F subpopulations would have different relative risks, so that the crude relative risk would be adjusted by weighing the relative risks of the subpopulations. In the format {crude, $\neg F$, F, adjusted} we get:

- Standardly, the epidemiologist would conclude to both interaction and confounding.

```
ETCAdjustedRRisk[CT[Data]] // N
```

```
{8.89314, 4.43333, 12.0098, 10.4945}
```

The point to observe however is that this is a different type of analysis than the one we are currently interested in. The crude and adjusted relative risks are measures for overall performance while our focus is on (i) determining which is the causal factor, or, if we know it, that we can explain why it is so, (ii) let us first get clarity on the influences of

the particular values before we worry about an "overall" measure. We can always weigh something but let us first get clarity on what we are weighing.

The model

The model consists of the 8 parameters that make the entire table. We can separate the total number *n,* and consider the remaining 7 probabilities. All these might be taken as constant and thus worthy of the label of being a parameter in the problem at hand. All this would be a happy state of affairs. What spoils this paradise are two snakes: (1) the proportions of the confounder may not be stable, (2) there are always the "other causes" ("causes not mentioned"). Given those snakes we want to make sure that our parameters are really constant. One way to do so is to use conditional probabilities.

The following definitions are useful, with *s, c* and *f* the marginal probabilities, *R* and *B* the influence of the cause on the success and *p* and *q* some important conditional probabilities. (Say, "*q*", from "quiet, not-confounding".)

Variable	*Meaning*	*Variable*	*Meaning*	*Variable*	*Meaning*
S	Success	s	$P[S]$	R	$P[S \mid C]$
C	Cause	c	$P[C]$	B	$P[S \mid \neg C]$
F	Confounder	f	$P[F]$	p	$P[C \mid F]$
				q	$P[C \mid \neg F]$

Note that generally $p > q$ since F will be a real confounder. Otherwise you would relabel the case so that bad weather instead of good weather becomes the confounder. It is less useful to consider $P[F \mid C]$ since then we would regard F and S as joint effects of C, while for a strong confounder we would not have the idea that it would depend upon C.

The "probability to get a success given the cause" is often called a "risk" since the success is often an outcome like a disease. If we take the perspective of the 2 × 2 basic world where we had only causes and effects, we would take $R = P[S \mid C]$ as the "parameter" that gives the size of the risk and $B = P[S \mid \neg C]$ as the "parameter" that gives the size of the background risk of "other causes". We would take them as constant and then regard them as the driving forces behind the whole process. We would compare R and B to the seeming risks from the confounder R_F and B_F. Note indeed that the marginal probability of the success can be decomposed as the background risk plus the average increment due to the cause, as holds in the 2 × 2 world and still holds in the 2 × 2 × 2 world.

$$s = P[S] = P[S, C] + P[S, \neg C] = R\,c + B\,(1 - c) = c\,(R - B) + B$$

$$s = P[S] = P[S, F] + P[S, \neg F] = R_F\,f + B_F\,(1 - f) = f\,(R_F - B_F) + B_F$$

ETCTable["ET", c, {R, B}]

	Cause	¬ Cause	Total
Success	$c\,R$	$B\,(1-c)$	$B\,(1-c)+c\,R$
¬ Success	$c\,(1-R)$	$(1-B)\,(1-c)$	$c\,B-B-c\,R+1$
Sum	c	$1-c$	1

However, in statistical terms R is an average. Of course still $R = \mathrm{P}[S \mid C] = \mathrm{P}[S, C] / \mathrm{P}[C]$ but now for both numerator and denominator we find a statistical dependence upon parameters p and q.

$$c = \mathrm{P}[C] = \mathrm{P}[C \mid F]\,f + \mathrm{P}[C \mid \neg F]\,(1 - f) = p\,f + q\,(1 - f)$$

$$\mathrm{P}[S, C] = \mathrm{P}[S, C, F] + \mathrm{P}[S, C, \neg F] = \mathrm{f}[p, q]$$

As a result of weighing by the probabilities p and q that are related to the confounder we find: (a) R and B are only averages and not constant over time, (b) there arise seeming risks so that the success seems to be related to the confounder. The "confusus magnitudinis" may come along with the "confusus directionis". Due to the confounder, we cannot just take the average, have to consider the presence of F and $\neg F$ as well, their relative proportions and their relation to the cause.

We can rewrite the three equations on s and c as $s = B + \beta\ c$, $s = B_F + \alpha\,f$ and $c = q + \gamma\,f$, eliminate c and then find:

$s = B + \beta\ c = B + \beta\ (q + \gamma\,f) = (B\ + \beta\,q) + (\beta\,\gamma)\,f = B_F + \alpha\,f$ so that $\alpha = \beta\,\gamma \Leftrightarrow B_F = B + q\,\beta$

It appears a pitfall to think that the lhs or rhs always hold. Just the equivalence holds. We wrote α, β and γ, but these actually are variates and not necessarily parameters. Only if $\alpha = \beta\,\gamma$ (under special conditions) then:

(i) $B_F = B + q\,(R - B)$

(ii) $(R_F - B_F) = (R - B)\,(p - q)$.

(iii) $R_F = B + q\,(R - B) + (R - B)\,(p - q) = B + p\,(R - B)$

The situation itself creates another possible confusion. Let us call it the "confusus contributionis". It may well be that the confounder has no direct influence on the disease, so that it is not a causal factor per se. But it may very well be that the confounder has a direct influence on the proportion of the population at risk. One might

call this a causal influence as well. Or not. It depends upon the situation and our state of confusion. For some events it may come as a surprise that we need to control for some characteristic and then that characteristic might be seen as a "causal factor". For some other events it might be obvious that it matters how the risk population is composed (e.g. males versus females) and then we might not think of this as "causal factor" but rather as a Simpson paradox.

The distribution of *Truth* and *Confounding* (summing over *Effect*) thus has *parameters* or *variables p* and *q* (parameters only when those are statistical regularities).

```
ETCTable["TC", f, {p, q}]
```

	Cause	¬ Cause	Total
Confounder	$f\ p$	$f\ (1-p)$	f
¬ Confounder	$(1-f)\ q$	$(1-f)\ (1-q)$	$1-f$
Sum	$f\ p+(1-f)\ q$	$-f\ p+f\ q-q+1$	1

Note that $p = q = c$ if and only if c and f are statistically independent (check the inner matrix). Since c is the independent factor, statistical independence means that we can substitute $p = q = c$. When there is no statistical independence then we may eliminate one variate, and the question arises whether this should be p, q or f. It appears most useful to eliminate p since we may take q and the absence of the confounder as the norm situation while it is useful to control f.

```
Solve[c == p f + q (1 - f), p]
```

$$\left\{\left\{p \to \frac{c+f\ q-q}{f}\right\}\right\}$$

From our definition of the case we must regard the distribution of truth, i.e. $\{c, 1 - c\}$ for $\{C, \neg C\}$, as the "driving" distribution, either from observation or controlled (influenced) by experiment. If there would be a causal relationship between cause and confounder (e.g. a common cause) so that p and q indeed are parameters then we would make another model. In the simplest case f has its own causes so that the relation between c and f is only "statistical". It would be too simple to assume that *Truth* and *Confounding* are distributed independently, and we allow for some statistical random effects. The distributions would rather be such that we might also think that the confounder is the cause, so that we really suffer the question whether we can determine conditions such that a seeming cause can be exposed as a confounder.

Notation in *Mathematica*

In the discussion below there will be computer output from routines in *Mathematica*. The following is a small legend for reading that output. The *R* and *B* are on the left hand side while *p* and *q* are on the right hand side. If the latter would deserve the name of being *parameters* (constant, unchanging) then one might consider that the Confounder caused the Cause. Yet, for now these are just statistical observations, just as the mentioned risks.

ETCPrTable[]

	Risk	Probability
1	ConditionalPr[Success][Cause]	ConditionalPr[Cause][Confounder]
2	ConditionalPr[Success][¬ Cause]	ConditionalPr[Cause][¬ Confounder]

Note that the above must be read as expressions for constants *f*[x_0 | y_0] and not as for variables *f*[*x* | *y*]. Since *Truth* ranges over {Cause, ¬Cause}, the above conditional probabilities don't depend upon *Truth* but upon its values Cause and ¬Cause.

Parameters from further conditionalizing

When we conditionalize further then we may find quantities that we might assume to be really constant so that they deserve the status of being a parameter.

The parameter of interest actually is r = P[S | C, ¬F], the conditional risk when the confounder is absent. Similarly, b = P[S | ¬C, ¬F] for the background risk. We will call a cause a "simple cause" when the effect only arises when the cause is present. A necessary condition is that $r = 1$ and $b = 0$.

Coming from a two-variable world we are confounded about the size of the effect when $R \neq r$ and $B \neq b$. Only under some conditions $s = r\,c + b\,(1 - c) = b + (r - b)\,c$.

This shows a crucial distinction between the ETC contingency table and other kinds of contingency tables. In other kinds of tables we have for example political preference for "Party A", "Party B" and Party C", and then the change from one cell to another does not necessarily have a causal connotation. For the ETC table the *absence* of the cause and / or the confounder have implications of huge importance, since they allow the direct identification of the individual effects. This identification of course is under the assumption that we know the truth, and in practice we have tables EXZ and EZX and have to compare them. But for now we just have one ETC. Thus, in terms of confouding, we now also have the "confusus causalitatis", holding that we might confuse an issue of mere association (any kind of contingency table) with an analysis of causality (the ETC

case). The Simpson paradox (**Appendix A**) works differently depending upon whether we do a causal analysis, in which the absence of the confounder has direct import, or whether we do an investigation into mere association.

We can also distinguish simple causality from more complex kinds of causality, where *C* for example is a mere contributing factor and not purely a simple cause. When we don't have simple causality then the presence of the cause needs to be qualified as to the presence of the confounder, and the absence of the cause (contributing factor) might still produce a small effect due to "other causes".

Before delving deeper into the formulas it will be useful to do the basic statistical analysis of the example contingency model, so that we can already recognize the variables that we have been introducing here: *n, s, c, f, p, q, r, b, R, B,* R_F and B_F.

The basic statistical analysis

The basic statistical analysis consists of identifying the proper conditional probabilities. In the following output, the first table give the true ratio table (discussed below). Subsequently there are three tables that give the simple border-matrices for two variables only. The first of these gives the relation between cause and effect that gives the average *R* and *B* from our two-variable world. The last table gives the seeming relation with seeming R_F and R_B that would arise if we would take the confouder as the cause.

- This takes the defaul CT[Data] derived from the table that has been set above. The routine also produces formal output, in a form that other routines can recognize. We can run another small routine to translate that output to human form. Note that you could easily suppress the formal output by putting a colon behind the call. Presently it is useful to show all output. There is also other output that we will discuss subsequently.

```
(res1 = Report[Example] = ETCStatistics[] // N) // MatrixForm
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	0.035	0.025	0.06
¬ Success	0.205	0.735	0.94
Sum	0.24	0.76	1.

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	82	11	93
¬ Success	374	533	907
Sum	456	544	1000

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	408	392	800
¬ Confounder	48	152	200
Sum	456	544	1000

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder	Total
Success	81	12	93
¬ Success	719	188	907
Sum	800	200	1000

$$\left(\begin{array}{l}
N \to 1000. \\
\text{NSuccess} \to 93. \\
\text{NCause} \to 456. \\
\text{NConfounder} \to 800. \\
\text{MarginalPr(Success)} \to 0.093 \\
\text{MarginalPr(Cause)} \to 0.456 \\
\text{MarginalPr(Confounder)} \to 0.8 \\
\text{IndependentPr(Truth, Confounding)} \to \text{False} \\
(\text{Success} \perp \neg\, \text{Confounder})(\text{Cause}) \to \text{False} \\
(\text{Success} \perp \neg\, \text{Confounder})(\neg\, \text{Cause}) \to \text{False} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}, \neg\, \text{Confounder}] \to 0.145833 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\, \text{Cause}, \neg\, \text{Confounder}] \to 0.0328947 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}, \text{Confounder}] \to 0.183824 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\, \text{Cause}, \text{Confounder}] \to 0.0153061 \\
\text{Risk} \to \begin{pmatrix} 0.183824 & 0.0153061 \\ 0.145833 & 0.0328947 \end{pmatrix} \\
\text{Interaction} \to \{\text{Add} \to 0.0555788,\ \text{Times} \to 7.57647\} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}] \to 0.179825 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\, \text{Cause}] \to 0.0202206 \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\text{Confounder}] \to 0.51 \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\neg\, \text{Confounder}] \to 0.24 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Confounder}] \to 0.10125 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\, \text{Confounder}] \to 0.06 \\
\text{RRisk(True)} \to 4.43333 \\
\text{RRisk(Cause)} \to 8.89314 \\
\text{RelativePr(Confounder)} \to 2.125 \\
\text{RRisk(Seeming)} \to 1.6875 \\
\text{ETCAdjustedRRisk} \to \{8.89314,\ 4.43333,\ 12.0098,\ 10.4945\} \\
\text{Conditions} \to \{\text{True},\ \text{True},\ \text{True},\ \text{True},\ \text{True}\} \\
\text{ConditionalPr}[\,\neg\, \text{Success}\,][\,\text{Cause}, \neg\, \text{Confounder}] \to 0.854167 \\
\text{ConditionalPr}[\,\neg\, \text{Success}\,][\,\neg\, \text{Cause}, \neg\, \text{Confounder}] \to 0.967105 \\
\text{ConditionalPr}[\,\neg\, \text{Success}\,][\,\text{Cause}, \text{Confounder}] \to 0.816176 \\
\text{ConditionalPr}[\,\neg\, \text{Success}\,][\,\neg\, \text{Cause}, \text{Confounder}] \to 0.984694 \\
\text{Safety} \to \begin{pmatrix} 0.816176 & 0.984694 \\ 0.854167 & 0.967105 \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} \text{False} & \text{False} \\ \text{False} & \text{False} \end{pmatrix} \\
\text{ETCSimpson} \to \{\text{Necessary} \to \text{False},\ \text{Sufficient} \to \{\text{True},\ \text{True},\ \text{False}\}\}
\end{array}\right)$$

Epidemiology concentrates on the relative risks. The true relative risk is 4.4 but due to the confounder the average relative risk is 8.9. If we would be confused about what would be the true cause then we would think that the average relative risk was 1.7. PM. Above we mentioned that we are less interested yet in such “overall measures” yet it is conventional to mention them, so we do here to.

```
ETCRiskTable[res1]
```

	Name	Value	Name	Value	Name	Value
Cause	r	0.145833	R	0.179825	Rf	0.10125
Background	b	0.0328947	B	0.0202206	Bf	0.06
Difference	r – b	0.112939	R – B	0.159604	Rf – Bf	0.04125
Ratio	r / b	4.43333	R / B	8.89314	Rf / Bf	1.6875

In the output of the ETCStatistics routine we find these conditions tested (see also the discussion below where the safety parameters are introduced):

(a) some (conditional) independences

(b) on risk: (1) $r > b$, (2) $R > B$, (3) $p \geq q$, (4) $R_F \geq B_F$, (5) $p / q \geq R_F / B_F$.

(c) on being a simple cause: {{$w = 0$, $v = 1$}, {$e = 1 - r = 0$ or $r = 1$, $a = 1 - b = 1$ or $b = 0$ }} (left column should be 0, right column should be 1).

(d) on the Simpson paradox: the necessary condition $b < r < 1 - v < 1 - w$ and the sufficient ones {$v > w$, $r > b$, $R < B$}.

In this discussion the seeming relative risk is important since we should allow that we don't have EXZ but EZX. The average outcome on the seeming risk may be a give-away. It might be confusing if we were to present both tables for EXZ and EZX in one presentation so that it is better to make separate runs (see below).

Regression coefficients

The identified risks can be compared to regression coefficients in a linear regression, since they express the contribution of a single causal event to the success. Yet, the discussion on regression coefficients is a bit more complex since events do not just come all by themselves, and are always classified in a table in more dimensions. The element-regression coefficients should also be distinguished from the vector-regression coefficients that arise when a unit of *Truth* is observed, as the vector {c, $1 - c$}. This is an issue to return to later. For now the following points are useful to keep in the back of your mind:

(1) When we have more observations, distinguished by time period (such as a year), then it might make sense not to aggregate all data as $s = \mathrm{P}[S] = c\,(R - B) + B = B + \beta\, c$ but to run a regression like $s_t = B + \beta\, c_t + \epsilon_t$ weighed by the numbers. This would give the weighed risk coefficients or the weighed conditional probabilities. That we can do this

confirms that we can understand the conditional probabilities as regression coefficients. The angle is important. (1a) We can further develop the regression model with more variables and parameters. (1b) We can locate an influence of time. Time is a "great confounder" and it would be usefully included in our analysis. (1c) Above regression is only on s and not on 1 - s. Above regression with c, R, B and the addition to 1 fully explains the situation, but including the error on 1 - s gives a different estimate when using weighted regression. If there are more values, also with confounding, it is useful to observe the inner elements of the tables and not just the margins, and to also use those observations. See the discussion on safety below.

(2) Of course the coefficients are unstable when we don't have the true model with the true fixed coefficients. This is why we would have this "confusus magnitudinis".

(3) Bayesians tend to think in terms of conditional probabilities and create the joint distributions from those. It may be that they just think in terms of regression coefficients. That method only holds when you use the proper fixed coefficients.

(4) For Ordinary Least Squares (OLS) the regression coefficient is related to the correlation coefficient by $\beta = \rho_{SC}\ \sigma_y / \sigma_x$. Taking just Bernouilli we might take $\sigma_y = \sqrt{s\,(1-s)}$ and $\sigma_x = \sqrt{c\,(1-c)}$ (or both equally affected by n), and then have our ρ_{SC}. Again, it would not be stable under confouding. And, again, this example would concern just a value of a variable and not the variable itself (we should be clear about what kind of correlation coefficient we want).

(5) The relation $c = f\,p + q\ (1 - f) = q + (p - q)\,f = q + \gamma\ f$ can also be seen as a regression, so that we can infer the implied correlation $\rho_{CF} = (p - q)\ \sigma_F / \sigma_C$. If $\rho_{CF} = 0$ then $p = q = c$ and the two distributions are independent. If $\rho_{CF} = 1$ then $\gamma = \sigma_C / \sigma_F$. This does not give any particular conclusion except that $q = c - \gamma\,f = c - \sigma_C / \sigma_F\,f = (c\ (1 - f) - \sigma_C\,\sigma_F) / (1 - f)$, which tells us how that particular cell $P[C, \neg F]$ looks like.

(6) For all equations $s = B + \beta\ c$, $s = B_F + \alpha\,f$ and $c = q + \gamma\,f$ the parameters are $\{\beta, \gamma, \alpha\}$, we might presume that these follow from regression and then we could infer the implied correlation matrix. This implied correlation matrix will be biased since these parameters need not be true parameters, actually may be variables, and one equation is fully dependent so that this is not a proper simultaneous equations model. But it can be interesting to have seen this, since it starts a line of reasoning.

- The equality test is on $\alpha = \beta\,\gamma$ and its consequence for B_F.

ETC222BiasedCorrelation[CT[Data]] // N

$$\left\{\text{Order} \to \{\text{Success, Cause, Confounder}\},\ \text{Mean} \to \{0.093,\ 0.456,\ 0.8\},\right.$$
$$\text{StandardDeviation} \to \{0.290432,\ 0.49806,\ 0.4\},\ \text{Coefficient} \to \{0.159604,\ 0.27,\ 0.04125\},$$
$$\left.\text{EqualityTests} \to \{\text{False, False}\},\ \text{Matrix} \to \begin{pmatrix} 1. & 0.273704 & 0.0568118 \\ 0.273704 & 1. & 0.216841 \\ 0.0568118 & 0.216841 & 1. \end{pmatrix}\right\}$$

Colignatus (2007g) develops the issue into a proper "risk difference regression" which uses all three variables and gives the following result. Note that ρ_{SF} switches sign. Apparently the direction of association between the confounder and the success is sensitive to whether the cause is present or not. Note that the equations contain interaction terms but these are not included in the correlation matrix.

RiskDiffRegress222[CT[Data], {S, C, F}, Spread → "Bernoulli"] // N

$$\left\{\text{Equations} \to \{S = 0.0555788\,F\,C + 0.112939\,C - 0.0175886\,F + 0.0328947,\right.$$
$$C = 0.0975344\,S\,F + 0.245058\,F + 0.365248\,S + 0.218085,$$
$$F = 0.203008\,S\,C + 0.166172\,C - 0.178748\,S + 0.724203\},$$
$$\text{CovarRegress} \to \begin{pmatrix} 0. & 0.112939 & -0.0175886 \\ 0.365248 & 0. & 0.245058 \\ -0.178748 & 0.166172 & 0. \end{pmatrix},\ \text{Method} \to \text{Bernoulli},$$
$$\text{FindMinimum} \to 0.00474879,\ \text{Spread} \to \{0.290432,\ 0.49806,\ 0.4\},$$
$$\left.\text{CorrelationMatrix} \to \begin{pmatrix} 1. & 0.197771 & -0.0669218 \\ 0.197771 & 1. & 0.182433 \\ -0.0669218 & 0.182433 & 1. \end{pmatrix}\right\}$$

Another measure of correlation can be based upon the "volume ratio", see Colignatus (2007d). Above regressions use marginals of the categories, while the volume ratio approach targets the variables and uses the inner matrices. When we run this routine then we find that the values of ρ_{SC} and ρ_{CF} are about the same as the biased correlation but ρ_{SF} changes sign as in the proper "risk difference regression". Colignatus (2007d & g) suggest that this would be the relevant correlation matrix while this present paper suggests that the relevant causal parameters are *r, b, w* and *v*.

NominalCorrelationMatrix[CT[Data]] // N

$$\begin{pmatrix} 1. & 0.26403 & -0.016659 \\ 0.26403 & 1. & 0.216104 \\ -0.016659 & 0.216104 & 1. \end{pmatrix}$$

(7) Let us stick to the correlations that are not determined by nominal correlations. Consider the choice of p and q. If we had observations over time, with table$_t$, c_t and f_t, then the best estimate might still come from using the table summed over time, yet the individual observations would still allow a better test of the question whether p and q would be constant. The issues are important since we must decide whether p and q are constant so that c follows from f, or that they are merely related, or whether we can control c and f with some p or q following. (Note that we currently don't model that F is a joint effect of C.) The above default assumption is that q is also controlled, so that we not only set c and f but also the proportion that the cause is present given that the confounder is absent. We should be able to do so in a (randomized) controlled trial. Conceivably, however, we have an observational study, and that proportion is decided for us. For repeated trials, a possible assumption is that ρ_{CF} is the constant parameter, so that when we select a value for c and one for f, then Nature chooses p and q from $\gamma = \rho_{CF}\ \sigma_C / \sigma_F$, $q = c - \gamma f$ and $p = q + \gamma$.

```
Thread[{p, q} == (ETCPQFromCorrelation[c, f, ρCF] // Simplify)]
```

$$\left\{p = c - (f-1)\sqrt{\frac{(f-1)f}{(c-1)c}}\ \rho_{CF},\ q = c - f\sqrt{\frac{(f-1)f}{(c-1)c}}\ \rho_{CF}\right\}$$

We cannot say anything about this in general, since everything depend upon the problem at hand.

The issue can be highlighted while using the figures from the example table, artificial as they are. Consider the average relative risk, and let us keep all other coefficients constant while we may vary over c, f and q. For this numerical example it appears that the issue is not too dramatic. The relative risks don't differ too much whatever we assume. Perhaps the cause and confounder are too close to independence.

- This gives $\rho_{CF} = (p - q)\ \sigma_F / \sigma_C = (c - q) / f\ \sigma_F / \sigma_C$.

```
impcorr = ETCImpliedCorrelation["NCause" / N,
    "NConfounder" / N, ConditionalPr["Cause"][! "Confounder"]] /. res1
```

0.216841

- This takes the numerical solution from the example contingency table, except for *c, f* and *q,* leaving them as parameters.

```
relrisk = R / B /. Thread[{R, B} → ETCAverageRisksFromSafety[
            {c, ConditionalPr["Success"]["Cause", ! "Confounder"],
              ConditionalPr["Success"][! "Cause", ! "Confounder"]},
            {f, q}, {ConditionalPr[! "Success"]["Cause", "Confounder"],
              ConditionalPr[! "Success"][! "Cause", "Confounder"]}
            ]] /.
  res1
```

$$\frac{(1-c)\left(0.183824-\frac{0.0379902\,(1-f)\,q}{c}\right)}{0.0328947\,(f-1)\,(q-1)+0.0153061\,(-c+f\,(1-q)+q)}$$

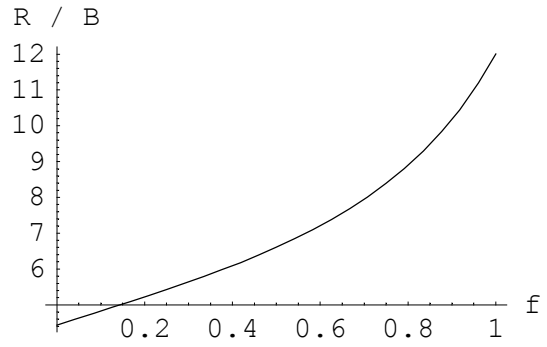
We now variate *f*, under three assumptions: (a) *p* and *q* are constant and *c* follows from *f,* (b) *c* and *q* are constant at the values in the example so that the adjustment is by *p*, (c) *c* is constant but *p* and *q* follow from constant correlation.

- (a) This is when *p* and *q* are constant. Note that the observed value is $f = 0.8$. The model is counter-intuitive, since it would presume that the confounder determines the presence of the cause.

```
relrisk /. {c → f p + q (1 - f)} /. {p → .51, q → .24} // Simplify
```

$$\frac{3.35714\,(f-2.81481)\,(f+0.595745)}{(f-1.42857)\,(f+0.888889)}$$

```
p1 = Plot[%, {f, 0, 1}, AxesLabel → {"f", "R / B"}];
```

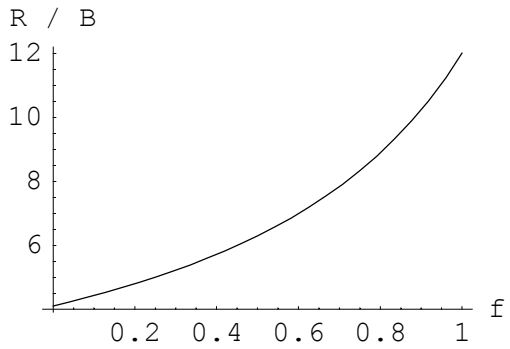


- (b) This is when we variate the confounder while keeping the cause at the same prevalence and keeping $q = P[C \mid \neg F]$ at a constant value. As (1 - *f*) drops, $p = P[C \mid F]$ must rise a bit. This would be the default assumption: that we can adjust *c* and *q* and *f,* so that both *p* and *R / B* would come out as the dependent variable. This would be typical of a randomized controlled trial.

```
relrisk /. {c → 0.456, q → 0.24}
```

$$\frac{0.544\,(0.183824 - 0.0199948\,(1 - f))}{0.0153061\,(0.76\,f - 0.216) - 0.025\,(f - 1)}$$

```
p2 = Plot[%, {f, 0, 1}, AxesLabel → {"f", "R / B"}];
```



- (c) This is when we observe variation in *c* and *f* while *p* and *q* are statistical regularities governed by the correlation between *C* and *F.* This might be typical of observational studies.

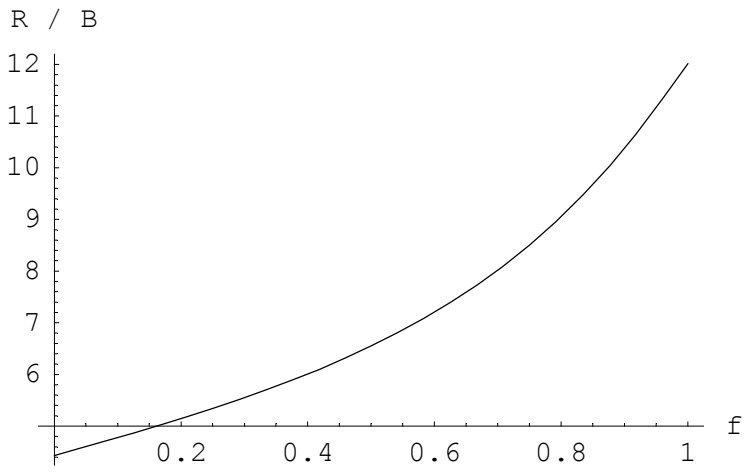
```
Thread[{p, q} → ETCFromCorrelation[0.456, f, impcorr]] // Simplify
```

$$\{p \to -0.435372\,\sqrt{-(f-1)\,f}\;f + 0.435372\,\sqrt{-(f-1)\,f} + 0.456,\; q \to 0.456 - 0.435372\,f\,\sqrt{-(f-1)\,f}\}$$

```
relrisk /. {c → f p + q (1 – f)} /.
    Thread[{p, q} → ETCFromCorrelation[0.456, f, impcorr]] // Simplify
```

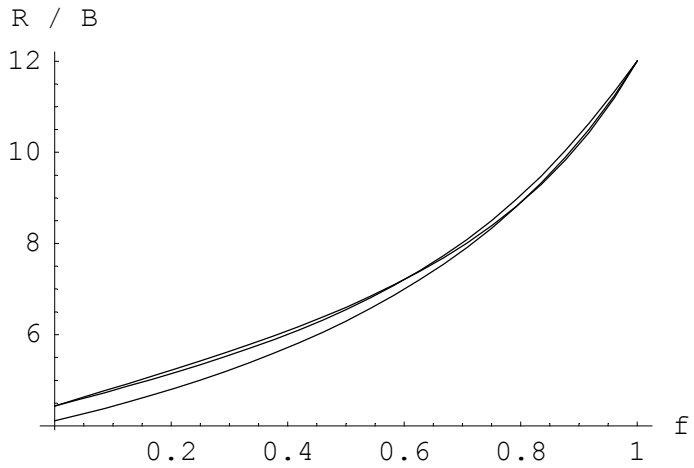
$$\frac{-0.0197318\,\sqrt{-(f-1)\,f}\;f^2 + 0.0197318\,\sqrt{-(f-1)\,f}\;f + 0.0206667\,f + 0.0793333}{-0.00765758\,\sqrt{-(f-1)\,f}\;f^2 + 0.00765758\,\sqrt{-(f-1)\,f}\;f - 0.00956821\,f + 0.0178947}$$

```
p3 = Plot[%, {f, 0, 1}, AxesLabel → {"f", "R / B"}];
```



- This combines all plots and shows that they are not too different, given this numerical example.

```
Show[p1, p2, p3];
```



(8) Below, we deduce that $B = b + q\,(r - b)$. One possible parameterization is to take B as given (as in the regression) and then deduce $q = (B - b) / (r - b)$. In a way, though, this merely shifts and possibly hides the problem. It does allow a quick link to the observed background risk yet obscures the link to the absence of the confounder and notably a possible relation between the absence of cause and absence of confounder. It would seem that it is better to be explicit about such a possible relation.

Other parameters that cause a success

Above, we identified r and b by means of conditioning. We first did the basic statistical analysis to show that this was a fruitful approach, and, to prevent that you were lost in the formulas. Now that we have seen some data and statistics, it will be a good moment to continue the formal analysis.

The data matrix contains two rows with a success, and we have only looked at one row, where the confounder is not present. When we take the row where the confounder is present as well then we might attribute the change of the risk just to that confounder.

- This takes the success part of the data matrix and conditions it. This is the small risk matrix that is printed in the output of the ETCStatistics routine. By symmetry, the background risk when F is present is the r_F if we would take the confounder as the real cause. The double-struck $\mathbb{r}$ gives the risk when both factors are present.

{{P[*S* | *C*, *F*], P[*S* | ¬ *C*, *F*]}, {P[*S* | *C*, ¬ *F*], P[*S* | ¬ *C*, ¬ *F*]}} == {{$\mathbb{r}$, r_F}, {r, b}}

$$\begin{pmatrix} P(S \mid C, F) & P(S \mid \neg\, C, F) \\ P(S \mid C, \neg\, F) & P(S \mid \neg\, C, \neg\, F) \end{pmatrix} = \begin{pmatrix} \mathbb{r} & r_F \\ r & b \end{pmatrix}$$

For the current model it would be strange that $\mathbb{r} \neq r$ or $r_F \neq b$. The idea is that F is a confounder and thus has no real influence. If the differences from r and b are importantly different from zero (and possibly even negative !) then a good reason needs to be given. One obvious explanation is that the success has not been measured correctly. But if the confounder e.g. is the distinction between Male and Female then one can imagine group specific risks.

What about the absence of the effect ?

Epidemiology focusses on the success, the disease. Yet, half of the cells are about the absence of the success. While writing this paper this author was for a long time entirely focussed on understanding and re-creating what the epidemiologists were doing, and thus focussed as well on the risks and not the safety. This tunnel vision might be called the “confusus focis ad risces”. Taking some distance from the risks, it appears that these safety parameters are important for a proper causal explanation.

- This takes the safety (or failure) part of the data matrix and conditionalizes it. This is the small safety matrix that is printed in the output of the ETCStatistics routine.

```
{{P[¬ S | C, F], P[¬ S | ¬ C, F]}, {P[¬ S | C, ¬ F], P[¬ S | ¬ C, ¬ F]}} ==
  {{w, v}, {e, a}}
```

$$\begin{pmatrix} P(\neg S \mid C, F) & P(\neg S \mid \neg C, F) \\ P(\neg S \mid C, \neg F) & P(\neg S \mid \neg C, \neg F) \end{pmatrix} = \begin{pmatrix} w & v \\ e & a \end{pmatrix}$$

Let u = P[¬C, ¬F] = P[¬S, ¬C, ¬F] + P[S, ¬C, ¬F] so that dividing both sides by u gives 1 = P[¬S | ¬C, ¬F] + P[S | ¬C, ¬F]. Hence a = 1 - b and in the same way we find that e = 1 - r.

What is important: (i) we now see that r and b have consequences with a safety interpretation, (ii) that v and w may be more sensible parameters than r_F and $\mathbb{r}$. Those may namely be related more to C and ¬C and less to this specific F.

- This gives the part for ¬S. Substraction from the total gives the part for S. Note that p can be eliminated as a function of c, f and q. In fact, we now have parameterized the whole contingency table (including the addition to 1).

```
ETCTable["–S|TC", {c, r, b}, {f, q}, {w, v}]
```

	Cause	¬ Cause
Confounder	$(c-(1-f)\,q)\,w$	$f\left(1-\frac{c-(1-f)\,q}{f}\right)v$
¬ Confounder	$(1-f)\,q\,(1-r)$	$(1-b)\,(1-f)\,(1-q)$
Sum	$(1-f)\,q\,(1-r)+(c-(1-f)\,q)\,w$	$(1-b)\,(1-f)\,(1-q)+f\left(1-\frac{c-(1-f)\,q}{f}\right.$

We can also deduce R and B, see **Appendix C** and **Appendix D**:

```
ETCAverageRisksFromSafety[{c, r, b}, {f, q}, {w, v}]
```

$$\left\{-w+\frac{(1-f)\,q\,(r+w-1)}{c}+1,\ \frac{b\,(f-1)\,(q-1)+(-c+f\,(1-q)+q)\,(1-v)}{1-c}\right\}$$

In the special case that $p = q = c$ then:

$$R = r\,(1-f)+(1-w)\,f$$

$$B = b\,(1-f)+(1-v)\,f$$

There are two cases to consider, identifiable as the columns in the small matrix of safety parameters:

(1) When the cause is not present then the situation should be totally safe (right column).

- If C is the only possible cause of S and it is a simple cause that always has effect then its absence should give full safety, $v = a = 1$.
- $a = P[\neg S \mid \neg C, \neg F] = 1 - b$ = background safety in total *absence*. If there is a background risk such that $b \neq 0$ then the background safety is reduced by the same amount.
- $v = P[\neg S \mid \neg C, F]$ = *safety* (using "*v*" from Dutch "veiligheid", the "*s*" already taken).
- Normally $v < a$ since including the confounder would reduce safety (otherwise possibly no reason to see it as a causal contender).
- If those equalities don't hold, then normally $v < 1$ and $a < 1$, and then there would be other factors that cause people to be less safe. If blocking C does not enhance full safety then one wouldn't call C a "simple cause" but rather a "contributing factor". To get a simple cause, we would redefine the absence of the cause as "absence of the original cause plus the presence of a truly effective block for other causes".

(2) The following are curious situations since the cause is present but there is no success (left column).

- If C is a simple cause of S then $w = e = 0$. Note that $e = 0$ means $r = 1$, another condition for a simple cause.
- $e = P[\neg S \mid C, \neg F]$ = exceptional (no confounder)
- $w = P[\neg S \mid C, F]$ = *miraculous* ("wunderbar", the "*m*" normally is an integer) (even the confounder present)
- Normally $w < e$ since including the confounder would reduce safety.
- Normally $w < v$ since including the cause would reduce safety (and it is miraculous when $w \neq 0$).
- If those equalities don't hold, and thus if the cause is present but the effect does not show, then something might actually be blocking the cause. For a "simple cause" we would redefine the cause as "unblocked cause", and recalculate the table. But if it concerns only a "contributing factor" then there is no miracle, since that concept allows that the cause does not always result into a success.

In **Appendix A** we derive the necessary conditions for the Simpson paradox that $b < r < 1 - v < 1 - w$, which translates too as $w < v < 1 - r$. For a causal process to get closer to the simple causal model we would require that $w \to 0$, $v \to 1$ and $r \to 1$. The causal

model requires that $v \to 1$ while the Simpson paradox requires that $v \to 0$. Under normal causal assumptions the Simpson paradox could not exist.

Reconstruction using safety

The former section mentioned that we have parameterized the whole $2 \times 2 \times 2$ matrix. See **Appendix C** for the actual deduction. The following shows how it works. From the 7 parameters and the summation to 1 the contingency table is created. It may be scaled up by multiplication with some n.

```
lis = SafetyToETCArray[{c, r, b}, {f, q}, {w, v}];

TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	$(c-(1-f)\,q)\,(1-w)$	$(-c+f\,(1-q)+q)\,(1-v)$
	¬ Confounder	$(1-f)\,q\,r$	$b\,(1-f)\,(1-q)$
¬ Success	Confounder	$(c-(1-f)\,q)\,w$	$(-c+f\,(1-q)+q)\,v$
	¬ Confounder	$(1-f)\,q\,(1-r)$	$(1-b)\,(1-f)\,(1-q)$

The probabilities of success are (transposing to keep the cause in the columns):

```
TableForm[lis[[1]] / (lis[[1]] + lis[[2]]) // Simplify // Transpose,
  TableHeadings → {{F, ¬ F}, {C, ¬ C}}]
```

	C	$\neg C$
F	$1-w$	$1-v$
$\neg F$	r	b

- Contingency tables actually only contain natural numbers but we have disabled a warning message on that. Note that in the output $r == R$ reads not as a declaration (where the LHS value is set) but as a condition that must evaluate to True or False. The output contains other conditions that we will discuss shortly.

```
(res = ETCStatistics[lis,  N → False] ) // Simplify // MatrixForm
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	$q\,r$	$b-b\,q$	$-q\,b+b+q\,r$
¬ Success	$q-q\,r$	$(b-1)\,(q-1)$	$b\,(q-1)-q\,r+1$
Sum	q	$1-q$	1

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause
Success	$-w\,c+c-(f-1)\,q\,(r+w-1)$	$b\,(f-1)\,(q-1)+(c+f\,(q-1)-q)\,(v-$
¬ Success	$c\,w+(f-1)\,q\,(r+w-1)$	$(-c+f-f\,q+q)\,v-(b-1)\,(f-1)\,(q-$
Sum	c	$1-c$

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	$c+(f-1)\,q$	$-c+f-f\,q+q$	f
¬ Confounder	$q-f\,q$	$(f-1)\,(q-1)$	$1-f$
Sum	c	$1-c$	1

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder
Success	$(c-q)\,(v-w)+f\,((q-1)\,v-q\,w+1)$	$(f-1)\,(b\,(q-1)-q\,r)$
¬ Success	$q\,(v-w)+c\,(w-v)+f\,(-q\,v+v+q\,w)$	$-(f-1)\,(b\,(q-1)-q\,r+1)$
Sum	f	$1-f$

$$\left\{\begin{array}{l}
N \to 1 \\
\text{NSuccess} \to b\,(f-1)\,(q-1)+q\,r+c\,v-q\,v-c\,w+q\,w-f\,(v+q\,(r-v+w)-1) \\
\text{NCause} \to c \\
\text{NConfounder} \to f \\
\text{IndependentPr(Truth, Confounding)} \to \frac{c-q}{f} = 0 \\
\text{(Success} \perp \neg \text{Confounder)(Cause)} \to \frac{(c+(f-1)\,q)\,(r+w-1)}{c} = 0 \\
\text{(Success} \perp \neg \text{Confounder)(}\neg \text{ Cause)} \to \frac{(c+f\,(q-1)-q)\,(b+v-1)}{c-1} = 0 \\
\text{ConditionalPr[Success][Cause, } \neg \text{ Confounder]} \to r \\
\text{ConditionalPr[Success][} \neg \text{ Cause, } \neg \text{ Confounder]} \to b \\
\text{ConditionalPr[Success][Cause, Confounder]} \to 1-w \\
\text{ConditionalPr[Success][} \neg \text{ Cause, Confounder]} \to 1-v \\
\text{Risk} \to \begin{pmatrix} 1-w & 1-v \\ r & b \end{pmatrix} \\
\text{Interaction} \to \{\text{Add} \to b-r+v-w,\ \text{Times} \to \frac{w-1}{v-1} - \frac{r}{b}\} \\
\text{ConditionalPr[Success][Cause]} \to \frac{-w\,c+c-(f-1)\,q\,(r+w-1)}{c} \\
\text{ConditionalPr[Success][} \neg \text{ Cause]} \to \frac{b\,(-q\,f+f+q-1)-(c+f\,(q-1)-q)\,(v-1)}{c-1} \\
\text{ConditionalPr[Cause][Confounder]} \to \frac{c+(f-1)\,q}{f} \\
\text{ConditionalPr[Cause][} \neg \text{ Confounder]} \to q \\
\text{ConditionalPr[Success][Confounder]} \to \frac{(c-q)\,(v-w)+f\,((q-1)\,v-q\,w+1)}{f} \\
\text{ConditionalPr[Success][} \neg \text{ Confounder]} \to -q\,b+b+q\,r \\
\text{RRisk(True)} \to \frac{r}{b} \\
\text{RRisk(Cause)} \to \frac{(c-1)\,(c\,(w-1)+(f-1)\,q\,(r+w-1))}{c\,(b\,(f-1)\,(q-1)+(c+f\,(q-1)-q)\,(v-1))} \\
\text{RelativePr(Confounder)} \to \frac{c+(f-1)\,q}{f\,q} \\
\text{RRisk(Seeming)} \to \frac{(c-q)\,(v-w)+f\,((q-1)\,v-q\,w+1)}{f\,(-q\,b+b+q\,r)} \\
\text{ETCAdjustedRRisk} \to \left\{\frac{(c-1)\,(c\,(w-1)+(f-1)\,q\,(r+w-1))}{c\,(b\,(f-1)\,(q-1)+(c+f\,(q-1)-q)\,(v-1))},\ \frac{r}{b},\ \frac{w-1}{v-1},\ \frac{r-f\,r}{b}+\frac{f\,(w-1)}{v-1}\right\} \\
\text{Conditions} \to \left\{r > b,\ \frac{-w\,c+c-(f-1)\,q\,(r+w-1)}{c} > \frac{b\,(-q\,f+f+q-1)-(c+f\,(q-1)-q)\,(v-1)}{c-1},\ \frac{c-q}{f} \geq 0,\ \frac{(c-q)\,(v-w)+f\,(b\,q}{f}\right. \\
\text{ConditionalPr[} \neg \text{ Success][Cause, } \neg \text{ Confounder]} \to 1-r \\
\text{ConditionalPr[} \neg \text{ Success][} \neg \text{ Cause, } \neg \text{ Confounder]} \to 1-b \\
\text{ConditionalPr[} \neg \text{ Success][Cause, Confounder]} \to w \\
\text{ConditionalPr[} \neg \text{ Success][} \neg \text{ Cause, Confounder]} \to v \\
\text{Safety} \to \begin{pmatrix} w & v \\ 1-r & 1-b \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} w=0 & v=1 \\ r=1 & b=0 \end{pmatrix} \\
\text{ETCSimpson} \to \left\{\text{Necessary} \to b < r < 1-v < 1-w,\ \text{Sufficient} \to \left\{v > w,\ r > b,\ \frac{b\,(-q\,f+f+q-1)-(c+f}{c-1}\right.\right.
\end{array}\right.$$

ETCRiskTable[res] // Rationalize // Simplify

	Name	Value	Name	Value
Cause	r	r	R	$\frac{-w c+c-(f-1) q (r+w-1)}{c}$
Background	b	b	B	$\frac{b (-q f+f+q-1)-(c+f (q-1)-q) (v-1)}{c-1}$
Difference	r – b	$r-b$	R – B	$\frac{b (f-1) (q-1)+(c+f (q-1)-q) (v-1)}{c-1}+\frac{-w c+c-(f-1) q (r+w-1)}{c}$
Ratio	r / b	$\frac{r}{b}$	R / B	$\frac{(c-1) (c (w-1)+(f-1) q (r+w-1))}{c (b (f-1) (q-1)+(c+f (q-1)-q) (v-1))}$

The seeming relative risk simplifies a bit when we eliminate *c*. This cannot be done in the routine since it does not know *p* as an independent variable.

psub = RRisk["Seeming"] /. res /. c → p f + (1 – f) q // Simplify

$$\frac{(p-1) v-p w+1}{-q b+b+q r}$$

When the ETC model is most powerful

Deliberately, we started with a matrix such that $r \neq 1$ and $b \neq 0$, since those are common applications. Yet, those situations also allow vagueness about the causality. The so-called cause then is actually a contributing factor only. The ETC model is most powerful when we consider a "simple cause" since then we can impose strong conditions on the parameters of the matrix. A cause is a simple cause when the success is recorded if and only if the cause has occurred. Discussions gain in clarity if causal chains can be broken down to those relations. Admittedly, models will always refer to "other causes" since it could well be impossible to exclude everything else. Probably the supreme counterfactual is to assume that there are no "other causes". Yet in prediction we often substitute $\epsilon = 0$ and then we eliminate those other causes. (Perhaps the human mind is continuously in the state of that supreme counterfactual, since modelling requires us to neglect things. The only thing that saves us is the ability to quickly switch to another model.)

- We also assume that cause and confounder are distributed independently.

lis = SafetyToETCArray[{c, 1, 0}, {f, c}, {0, 1}];

TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]

		Cause	¬ Cause
Success	Confounder	$c - c\,(1 - f)$	0
	¬ Confounder	$c\,(1 - f)$	0
¬ Success	Confounder	0	$(1 - c)\,f$
	¬ Confounder	0	$(1 - c)\,(1 - f)$

- The confounder is exposed by having a relative risk of 1.

(res = ETCStatistics[lis, N → False]) // Simplify // MatrixForm

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	c	0	c
¬ Success	0	$1 - c$	$1 - c$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	c	0	c
¬ Success	0	$1 - c$	$1 - c$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	$c\,f$	$f - c\,f$	f
¬ Confounder	$c - c\,f$	$(c - 1)\,(f - 1)$	$1 - f$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder	Total
Success	$c\,f$	$c - c\,f$	c
¬ Success	$f - c\,f$	$(c - 1)\,(f - 1)$	$1 - c$
Sum	f	$1 - f$	1

$$\left(\begin{array}{l}
N \to 1 \\
\text{NSuccess} \to c \\
\text{NCause} \to c \\
\text{NConfounder} \to f \\
\text{IndependentPr(Truth, Confounding)} \to \text{True} \\
(\text{Success} \perp \neg\ \text{Confounder})(\text{Cause}) \to \text{True} \\
(\text{Success} \perp \neg\ \text{Confounder})(\neg\ \text{Cause}) \to \text{True} \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \text{Cause}, \neg\ \text{Confounder}] \to 1 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \neg\ \text{Cause}, \neg\ \text{Confounder}] \to 0 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \text{Cause}, \text{Confounder}] \to 1 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \neg\ \text{Cause}, \text{Confounder}] \to 0 \\
\text{Risk} \to \begin{pmatrix} 1 & 0 \\ 1 & 0 \end{pmatrix} \\
\text{Interaction} \to \{\text{Add} \to 0, \text{Times} \to \text{Indeterminate}\} \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \text{Cause}] \to 1 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \neg\ \text{Cause}] \to 0 \\
\text{ConditionalPr}[\ \text{Cause}\ ][\ \text{Confounder}] \to c \\
\text{ConditionalPr}[\ \text{Cause}\ ][\ \neg\ \text{Confounder}] \to c \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \text{Confounder}] \to c \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \neg\ \text{Confounder}] \to c \\
\text{RRisk(True)} \to \infty \\
\text{RRisk(Cause)} \to \infty \\
\text{RelativePr(Confounder)} \to 1 \\
\text{RRisk(Seeming)} \to 1 \\
\text{ETCAdjustedRRisk} \to \{\infty, \infty, \infty, \infty\} \\
\text{Conditions} \to \{\text{True}, \text{True}, \text{True}, \text{True}, \text{True}\} \\
\text{ConditionalPr}[\ \neg\ \text{Success}\ ][\ \text{Cause}, \neg\ \text{Confounder}] \to 0 \\
\text{ConditionalPr}[\ \neg\ \text{Success}\ ][\ \neg\ \text{Cause}, \neg\ \text{Confounder}] \to 1 \\
\text{ConditionalPr}[\ \neg\ \text{Success}\ ][\ \text{Cause}, \text{Confounder}] \to 0 \\
\text{ConditionalPr}[\ \neg\ \text{Success}\ ][\ \neg\ \text{Cause}, \text{Confounder}] \to 1 \\
\text{Safety} \to \begin{pmatrix} 0 & 1 \\ 0 & 1 \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} \text{True} & \text{True} \\ \text{True} & \text{True} \end{pmatrix} \\
\text{ETCSimpson} \to \{\text{Necessary} \to \text{False}, \text{Sufficient} \to \{\text{True}, \text{True}, \text{False}\}\}
\end{array}\right)$$

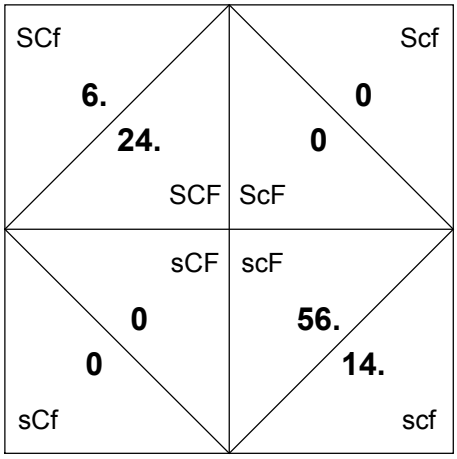
In this case the standard concept of relative risk does not find useful employment.

ETCRiskTable[res] // Rationalize // Simplify

	Name	Value	Name	Value	Name	Value
Cause	r	1	R	1	Rf	*c*
Background	b	0	B	0	Bf	*c*
Difference	r − b	1	R − B	1	Rf − Bf	0
Ratio	r / b	ComplexInfinity	R / B	ComplexInfinity	Rf / Bf	1

Selecting some parameter values shows us this layout in the ETC square: a diagonal of numbers and a cross-diagonal of zeros.

```
ETCSquare[lis 100 /. {c → 0.3, f → 0.8}];
```



Formal analysis on the risk approach

Above we took the safety parameters as the keys for reconstructing the contingency table. We might also focus on risk and perform a different parameterization. When writing this article, this was actually the first result. But the approach with the safety parameters is most insightful and deserved the top position in the discussion above. Now, however, it is proper to also consider the risk parameterization.

Instead of the averages we are interested in the driving risks (using above TC table):

$$r = \mathrm{P}[S \mid C, \neg F] = \mathrm{P}[S, C, \neg F] / (\mathrm{P}[C \mid \neg F]\, \mathrm{P}[\neg F]) = \mathrm{P}[S, C, \neg F] / (q\,(1 - f))$$

$$b = \mathrm{P}[S \mid \neg C, \neg F] = \mathrm{P}[S, \neg C, \neg F] / (\mathrm{P}[\neg C \mid \neg F]\, \mathrm{P}[\neg F]) = \mathrm{P}[S, \neg C, \neg F] / ((1 - q)\,(1 - f))$$

And this allows us to understand what happens in general when the confounder is not present:

$$\mathrm{P}[S, C \mid \neg F] = \mathrm{P}[S, C, \neg F] / \mathrm{P}[\neg F] = r\, q$$

$$\mathrm{P}[S, \neg C \mid \neg F] = \mathrm{P}[S, \neg C, \neg F] / \mathrm{P}[\neg F] = b\,(1 - q)$$

When we consider the group $\neg F$ as a whole, conditionally, then we find the following table - which is also the second table printed in the above ETCStatistics output (i.e. the Ratio table).

- This looks only at the group with $\neg F$. All values must be multiplied by 1 - *f*.

```
ETCTable["ET", q, {r, b}]
```

	Cause	¬ Cause	Total
Success	$q\,r$	$b\,(1-q)$	$b\,(1-q)+q\,r$
¬ Success	$q\,(1-r)$	$(1-b)\,(1-q)$	$q\,b-b-q\,r+1$
Sum	q	$1-q$	1

If we multiply this with 1 - *f* and substract this result from the earlier total, we get the matrix for the group with *F*. Hence we have parameterized the whole $2 \times 2 \times 2$ matrix again. See **Appendix D** for a review. In this parameterization *R* and *B* are retained as input parameters even though they are the outcome of the causal process. It can be useful to have this flexibility for creating contingency tables.

Reconstruction using average risks

The following creates a contingency table from the average risks. From the 7 parameters and the summation to 1 the matrix is created. It may be scaled up by multiplication with some *n*. When the matrix is submitted to the routine for the statistics then we get the proper results that fit the earlier tables.

- Note that we include *R* and *B* as variable. In empirical observation these are outcomes but table designers like us will want to control how the table will look like.

```
lis = RiskToETCArray[{c, r, b}, {f, q}, {R, B}];

TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	$c\,R-(1-f)\,q\,r$	$B\,(1-c)-b\,(1-f)\,(1-q)$
	¬ Confounder	$(1-f)\,q\,r$	$b\,(1-f)\,(1-q)$
¬ Success	Confounder	$c\,(1-R)-(1-f)\,q\,(1-r)$	$(1-B)\,(1-c)-(1-b)\,(1-f)\,(1$
	¬ Confounder	$(1-f)\,q\,(1-r)$	$(1-b)\,(1-f)\,(1-q)$

- The statistics routine takes the matrix as it is and its origin does not matter. The advantage of the risk parameterization is that we now recognize more output. For example, *s* was unrecognizable before but now simplifies to *s* = *B* + *c* (*R* - *B*).

```
(res = ETCStatistics[lis, N → False]) // Simplify // MatrixForm
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	$q\,r$	$b - b\,q$	$-q\,b + b + q\,r$
¬ Success	$q - q\,r$	$(b-1)\,(q-1)$	$b\,(q-1) - q\,r + 1$
Sum	q	$1 - q$	1

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	$c\,R$	$B - B\,c$	$-c\,B + B + c\,R$
¬ Success	$c - c\,R$	$(B-1)\,(c-1)$	$B\,(c-1) - c\,R + 1$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	$c + (f-1)\,q$	$-c + f - f\,q + q$	f
¬ Confounder	$q - f\,q$	$(f-1)\,(q-1)$	$1 - f$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder
Success	$-c\,B + B + b\,(-q\,f + f + q - 1) + f\,q\,r - q\,r + c\,R$	$(f-1)\,(b\,(q-1) - q\,r$
¬ Success	$B\,(c-1) + f + b\,(f-1)\,(q-1) - f\,q\,r + q\,r - c\,R$	$-(f-1)\,(b\,(q-1) - q$
Sum	f	$1 - f$

$$\left\{
\begin{array}{l}
N \to 1 \\
\text{NSuccess} \to -c\,B + B + c\,R \\
\text{NCause} \to c \\
\text{NConfounder} \to f \\
\text{IndependentPr(Truth, Confounding)} \to \frac{c-q}{f} = 0 \\
(\text{Success} \perp \neg\ \text{Confounder})(\text{Cause}) \to r = R \\
(\text{Success} \perp \neg\ \text{Confounder})(\neg\ \text{Cause}) \to b = B \\
\text{ConditionalPr}[\text{ Success }][\text{ Cause, } \neg\ \text{Confounder}] \to r \\
\text{ConditionalPr}[\text{ Success }][\neg\ \text{Cause, } \neg\ \text{Confounder}] \to b \\
\text{ConditionalPr}[\text{ Success }][\text{ Cause, Confounder}] \to \frac{(f-1)\,q\,r+c\,R}{c+(f-1)\,q} \\
\text{ConditionalPr}[\text{ Success }][\neg\ \text{Cause, Confounder}] \to \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q} \\
\text{Risk} \to \begin{pmatrix} \frac{(f-1)\,q\,r+c\,R}{c+(f-1)\,q} & \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q} \\ r & b \end{pmatrix} \\
\text{Interaction} \to \left\{\text{Add} \to b - r + \frac{(f-1)\,q\,r+c\,R}{c+(f-1)\,q} - \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q},\ \text{Times} \to \frac{(c+f\,(q-1)-q)\,((f-1)\,q\,r+c\,R)}{(B\,(c-1)+b\,(f-1)\,(q-1))\,(c+(f-1)\,q)}\right. \\
\text{ConditionalPr}[\text{ Success }][\text{ Cause}] \to R \\
\text{ConditionalPr}[\text{ Success }][\neg\ \text{Cause}] \to B \\
\text{ConditionalPr}[\text{ Cause }][\text{ Confounder}] \to \frac{c+(f-1)\,q}{f} \\
\text{ConditionalPr}[\text{ Cause }][\neg\ \text{Confounder}] \to q \\
\text{ConditionalPr}[\text{ Success }][\text{ Confounder}] \to \frac{-c\,B+B+b\,(-q\,f+f+q-1)+f\,q\,r-q\,r+c\,R}{f} \\
\text{ConditionalPr}[\text{ Success }][\neg\ \text{Confounder}] \to -q\,b + b + q\,r \\
\text{RRisk(True)} \to \frac{r}{b} \\
\text{RRisk(Cause)} \to \frac{R}{B} \\
\text{RelativePr(Confounder)} \to \frac{c+(f-1)\,q}{f\,q} \\
\text{RRisk(Seeming)} \to \frac{-c\,B+B+b\,(-q\,f+f+q-1)+f\,q\,r-q\,r+c\,R}{f\,(-q\,b+b+q\,r)} \\
\text{ETCAdjustedRRisk} \to \left\{\frac{R}{B},\ \frac{r}{b},\ \frac{(c+f\,(q-1)-q)\,((f-1)\,q\,r+c\,R)}{(B\,(c-1)+b\,(f-1)\,(q-1))\,(c+(f-1)\,q)},\ \frac{r-f\,r}{b} + \frac{f\,(c+f\,(q-1)-q)\,((f-1)\,q\,r+c\,R)}{(B\,(c-1)+b\,(f-1)\,(q-1))\,(c+(f-1)\,q)}\right\} \\
\text{Conditions} \to \left\{r > b,\ R > B,\ \frac{c-q}{f} \geq 0,\ \frac{-c\,B+B+b\,(q-1)-q\,r+c\,R}{f} \geq 0,\ \frac{b\,c\,(q-1)+q\,(-c\,B+B+c\,(R-r))}{f\,q\,(b\,(q-1)-q\,r)} \geq 0\right\} \\
\text{ConditionalPr}[\neg\ \text{Success }][\text{ Cause, } \neg\ \text{Confounder}] \to 1 - r \\
\text{ConditionalPr}[\neg\ \text{Success }][\neg\ \text{Cause, } \neg\ \text{Confounder}] \to 1 - b \\
\text{ConditionalPr}[\neg\ \text{Success }][\text{ Cause, Confounder}] \to \frac{-R\,c+c+q\,(-r\,f+f+r-1)}{c+(f-1)\,q} \\
\text{ConditionalPr}[\neg\ \text{Success }][\neg\ \text{Cause, Confounder}] \to \frac{-c\,B+B+c-f+f\,q-q+b\,(-q\,f+f+q-1)}{c+f\,(q-1)-q} \\
\text{Safety} \to \begin{pmatrix} \frac{-R\,c+c+q\,(-r\,f+f+r-1)}{c+(f-1)\,q} & \frac{-c\,B+B+c-f+f\,q-q+b\,(-q\,f+f+q-1)}{c+f\,(q-1)-q} \\ 1-r & 1-b \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} \frac{(f-1)\,q\,(r-1)+c\,(R-1)}{c+(f-1)\,q} = 0 & \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q} = 0 \\ r = 1 & b = 0 \end{pmatrix} \\
\text{ETCSimpson} \to \left\{\text{Necessary} \to b < r < \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q} < \frac{(f-1)\,q\,r+c\,R}{c+(f-1)\,q},\ \text{Sufficient} \to \left\{\frac{-c\,B+B+c-f+f\,q}{c+f}\right.\right.
\end{array}
\right.$$

ETCRiskTable[res] // Rationalize // Simplify

	Name	Value	Name	Value	Name	Value
Cause	r	r	R	R	Rf	$\frac{-c\,B+B+b\,(-q\,f+f+q-1)+f\,q\,r-q\,r+c}{f}$
Background	b	b	B	B	Bf	$-q\,b+b+q\,r$
Difference	r – b	$r-b$	R – B	$R-B$	Rf – Bf	$\frac{-c\,B+B+b\,(q-1)-q\,r+c\,R}{f}$
Ratio	r / b	$\frac{r}{b}$	R / B	$\frac{R}{B}$	Rf / Bf	$\frac{-c\,B+B+b\,(-q\,f+f+q-1)+f\,q\,r-q\,r+c}{f\,(-q\,b+b+q\,r)}$

The seeming relative risk does not simplify when we eliminate *c*.

psub = RRisk["Seeming"] /. res /. c → p f + q (1 – f) // Simplify

$$\frac{b\,(f-1)\,(q-1)+B\,(f\,p-f\,q+q-1)-f\,q\,r+q\,r-f\,p\,R+f\,q\,R-q\,R}{f\,(b\,(q-1)-q\,r)}$$

PM. If we calculate *R* and *B* from the safety parameters and use these to create the table again, then we get the same results. See **Appendix D**.

Intermediate conclusions

It will help to summarize what we have done up to now so that we can use that as a base for the subsequent discussion.

1. We designed a statistics routine that analyzes a count data contingency table into the various marginal and conditional probabilities.
2. We identified the proper risk and safety parameters, as opposed to statistical averages.
3. We designed a routine to create a contingency table with count data, by reverse-engineering from true parameters.
4. We designed another routine to reverse-engineer but using average risks.
5. We identified conditions $\{r, b\} = \{1, 0\}$ and $\{w, v\} = \{0, 1\}$ for when the cause can be called a "simple cause" (a success if and only if a cause). These values imply relative freedom or conditional independence, but not conversely. If those values are not present then the cause is not a simple cause anymore, just a "contributing factor".
6. We identified interdepencies between parameters and variables: (6a) *p* depends upon other more useful parameters, (6b) we cannot set the safety parameters and the averages at the same time.

7. We identified when the distributions of cause and confounder would be statistically independent ($p = q = c$).

For the following, we will delve deeper into the issue of statistical independence. It appears that epidemiology uses this standardly as a frame of reference. Epidemiologists also focus on the relation between $\{r, b\}$ and $\{R, B\}$. To link up to the literature it will be useful to consider independence while using the parameterization of RiskToETCArray instead of SafetyToETCArray.

Considering the case when $p = q = c$

To what extent are p and q really "parameters" ? Or, to what extent are they merely the product of sampling in a perhaps random reality ? There are two key solution approaches:

- The cause and confounder are statistically independent. Then $p = q = c$.
- The cause and confounder are statistically dependent. It may just be the case that the numbers suggest a relation even though there isn't one (since we are discussing a *true* confouder). It p and q have stable values and indeed are parametric rather than f itself, then $f = (c - q) / (p - q)$. (PM. If one substitutes $p = q$ then they are equal to c again; then f seems indeterminate, but it isn't, since f is given from the marginal. Only if $p \neq q$ and if we regard them as parameters then and only then we solve $f = (c - q) / (p - q)$.)

The general approach thus is to allow for both dependence and independence, where the researcher must provide a statistical explanation when the variables and the distributions $\{c, 1 - c\}$ and $\{f, 1 - f\}$ are dependent.

The following sets $p = q = c$.

```
lis = RiskToETCArray[{c, r, b}, {f, c}, {R, B}];
```

```
TableForm[lis, TableHeadings → CT["ETC", TableHeadings]] // Simplify
```

		Cause	¬ Cause
Success	Confounder	$c((f-1)r+R)$	$-(c-1)(B+b(f-1))$
	¬ Confounder	$-c(f-1)r$	$b(c-1)(f-1)$
¬ Success	Confounder	$c(-rf+f+r-R)$	$(c-1)(B+b(f-1)-f)$
	¬ Confounder	$c(f-1)(r-1)$	$-(b-1)(c-1)(f-1)$

```
(res = ETCStatistics[lis, N → False] // Simplify) // MatrixForm
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	$c\,r$	$b - b\,c$	$-c\,b + b + c\,r$
¬ Success	$c - c\,r$	$(b-1)\,(c-1)$	$b\,(c-1) - c\,r + 1$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	$c\,R$	$B - B\,c$	$-c\,B + B + c\,R$
¬ Success	$c - c\,R$	$(B-1)\,(c-1)$	$B\,(c-1) - c\,R + 1$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	$c\,f$	$f - c\,f$	f
¬ Confounder	$c - c\,f$	$(c-1)\,(f-1)$	$1 - f$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder
Success	$-c\,B + B + b\,(-f\,c + c + f - 1) + c\,((f-1)\,r + R)$	$(f-1)\,(b\,(c-1) - c\,r)$
¬ Success	$B\,(c-1) + b\,(f-1)\,(c-1) + f + c\,r - c\,f\,r - c\,R$	$-(f-1)\,(b\,(c-1) - c$
Sum	f	$1 - f$

$$
\begin{cases}
N \to 1 \\
\text{NSuccess} \to -c\,B + B + c\,R \\
\text{NCause} \to c \\
\text{NConfounder} \to f \\
\text{IndependentPr(Truth, Confounding)} \to \text{True} \\
(\text{Success} \perp \neg\,\text{Confounder})(\text{Cause}) \to r = R \\
(\text{Success} \perp \neg\,\text{Confounder})(\neg\,\text{Cause}) \to b = B \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}, \neg\,\text{Confounder}] \to r \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause}, \neg\,\text{Confounder}] \to b \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}, \text{Confounder}] \to \frac{(f-1)\,r+R}{f} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause}, \text{Confounder}] \to \frac{B+b\,(f-1)}{f} \\
\text{Risk} \to \begin{pmatrix} \frac{(f-1)\,r+R}{f} & \frac{B+b\,(f-1)}{f} \\ r & b \end{pmatrix} \\
\text{Interaction} \to \left\{\text{Add} \to \frac{b-B-r+R}{f},\ \text{Times} \to \frac{b\,R-B\,r}{b\,(B+b\,(f-1))}\right\} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}] \to R \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause}] \to B \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\text{Confounder}] \to c \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\neg\,\text{Confounder}] \to c \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Confounder}] \to \frac{-c\,B+B+b\,(-f\,c+c+f-1)+c\,((f-1)\,r+R)}{f} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Confounder}] \to -c\,b + b + c\,r \\
\text{RRisk(True)} \to \frac{r}{b} \\
\text{RRisk(Cause)} \to \frac{R}{B} \\
\text{RelativePr(Confounder)} \to 1 \\
\text{RRisk(Seeming)} \to \frac{-c\,B+B+b\,(-f\,c+c+f-1)+c\,((f-1)\,r+R)}{f\,(-c\,b+b+c\,r)} \\
\text{ETCAdjustedRRisk} \to \left\{\frac{R}{B},\ \frac{r}{b},\ \frac{(f-1)\,r+R}{B+b\,(f-1)},\ \frac{b\,((f-1)\,r+f\,R)-B\,(f-1)\,r}{b\,(B+b\,(f-1))}\right\} \\
\text{Conditions} \to \left\{r > b,\ R > B,\ \text{True},\ \frac{-c\,B+B+b\,(c-1)+c\,(R-r)}{f} \geq 0,\ \frac{-c\,b+b+B\,(c-1)+c\,(r-R)}{f\,(-c\,b+b+c\,r)} \geq 0\right\} \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\text{Cause}, \neg\,\text{Confounder}] \to 1 - r \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\neg\,\text{Cause}, \neg\,\text{Confounder}] \to 1 - b \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\text{Cause}, \text{Confounder}] \to \frac{-r\,f+f+r-R}{f} \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\neg\,\text{Cause}, \text{Confounder}] \to \frac{-f\,b+b-B+f}{f} \\
\text{Safety} \to \begin{pmatrix} \frac{-r\,f+f+r-R}{f} & \frac{-f\,b+b-B+f}{f} \\ 1-r & 1-b \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} \frac{f\,(r-1)-r+R}{f} = 0 & \frac{B+b\,(f-1)}{f} = 0 \\ r = 1 & b = 0 \end{pmatrix} \\
\text{ETCSimpson} \to \left\{\text{Necessary} \to b < r < \frac{B+b\,(f-1)}{f} < \frac{(f-1)\,r+R}{f},\ \text{Sufficient} \to \left\{\frac{-f\,b+b-B+f\,r-r+R}{f} > 0,\ r >\right.\right.
\end{cases}
$$

The difference with the former result on risk is in the last column.

ETCRiskTable[res] // Rationalize // Simplify

	Name	Value	Name	Value	Name	Value
Cause	r	r	R	R	Rf	$\frac{-c\,B+B+b\,(-f\,c+c+f-1)+c\,((f-1)\,r+}{f}$
Background	b	b	B	B	Bf	$-c\,b+b+c\,r$
Difference	r − b	$r-b$	R − B	$R-B$	Rf − Bf	$\frac{-c\,B+B+b\,(c-1)+c\,(R-r)}{f}$
Ratio	r / b	$\frac{r}{b}$	R / B	$\frac{R}{B}$	Rf / Bf	$\frac{-c\,B+B+b\,(-f\,c+c+f-1)+c\,((f-1)\,r+}{f\,(-c\,b+b+c\,r)}$

The seeming relative risk R_F / B_F does not simplify, and becomes more complex if we would substitute *c*.

RRisk["Seeming"] /. res // Simplify

$$\frac{-c\,B+B+b\,(-f\,c+c+f-1)+c\,((f-1)\,r+R)}{f\,(-c\,b+b+c\,r)}$$

Conditional independence or relative freedom

In our two-variable world we had $Y = S$ as the variable to be explained and explanatory variable $X = C$. Now a new variable $Z = F$ is added.

If P[Y | X, Z] = P[Y | X] then explanatory variable X *c*ontains all information and is sufficient for the conditional probability between Y and X. Then Y and Z are said to be "conditionally independent" given X, and this is denoted as ($Y \perp Z$ | X). A shorter and clearer English expression is that Y and Z are "free from each other" relative to X.

- This denotes the logical statement ($Y \perp Z$ | X), but "|" in *Mathematica* stands for the Alternatives pattern, which gives problems in replacement. This is also the reason why we use ConditionalPr[...][...] instead of the bar.

  FreePr[Y, Z][X]

  $(Y \perp Z)(X)$

Note that there is a difference between such a relation for some fixed constants like {X_0, X_1} and the variables X that take those values. For the relation to hold for variables it must hold for all their values.

It would be useful to define as well that P[Y || X] = $\forall\, Z : (Y \perp Z \mid X)$. As Hintikka remarked, a quantifier always has some domain, and this quantifier runs over the available concepts in the domain of discussion, either the variables or their values, so that if X is a variable then Z too. The double bar expresses that X is necessary and

sufficient for Y and that there cannot be any confounding (for all variables that are available in the domain of discussion).

Schield (2003) remarks: "Students often think of numerical associations as immutable—as unconditional. By studying Simpson's Paradox students overcome this mistaken perception." Those students not only confuse P[X, Y] and P[$Y \mid X$] but thus also P[$Y \mid X$] with P[$Y \mid\mid X$]. Let us call this the "confusus libertatis".

PM 1. This also causes the thought that advances in mathematical notation are made by catching confusions by students.

PM 2. Conditional independence is often presented as: "If P[$X \mid Y, Z$] = P[$X \mid$ Z] then X and Y are conditionally independent given Z, and this is denoted as ($X \perp Y \mid Z$)." This presentation derives from the alphabetical order X, Y, Z, and it derives from the didactics of teachers in statistics who want to have their alphabet in neat order. But, in proper didactics, the focus of the student is on X and Y, and not on X and Z. The student has been working in the two-variable world and suddenly there is a third. The normal risk-averse student will tend to regard this Z as less relevant, and neglect it, like an austrich will hide its head in the sand or like children hide behind their hands or under a blanket. Only the minority of risk-prone students will focus on this new event Z and be willing to accept that it suddenly is the more important variable in a new definition. We better serve the majority, the risk-averse students. We also had the convention that X is the explanatory variable so that ($S \perp F \mid C$) = ($Y \perp Z \mid X$) anyway. We should not require students to suddenly invert all variable relationships, just to get a neat alphabetical order. Possibly it are not the students but the teachers who are a bit confused, which could be the "confusus doctoris".

PM 3. We already mentioned the layout of Kleinbaum et al. (2003), with the issue of confounding presented in chapter 10 and not in chapter 2. The current chapter 2 gives an overview of epidemiology. But the title of the book and CD is "ActivEpi". The overview confronts the student with all kinds of concepts that cannot be actively applied since they are only half understood. The long list of new topics might be called technically an "overview" but the real meaning of "overview" is to create insight. This would be another case of "confusus doctoris". It would be much better to actually start doing a case, in "learning by doing" and "hands on studying", where both 2 × 2 and 2 × 2 × 2 tables are used, such that the 2 × 2 × 2 table helps to understand what the 2 × 2 table means. Once the student has mastered the ETC format and has a sense of accomplishment then one can proceed with calculating incidence rates, relating counts to

person years, to show that epidemiology is more complicated. PM. That book would also benefit from printing in columns, given its pagewidth.

PM 4. It is advisable to use the phrase "relative freedom" as equivalent to "conditional independence". The latter phrase is a technical term from the realm of theorists who work with the concept on a daily basis. The normal student will be put off, however, and hide under the blanket again. Independence is like freedom however and conditioning in probability theory is just seeing probabilities in their relative proportions. The term "relative freedom" is more student-friendly.

PM 5. Students confuse P[$X \mid Y$] and P[$Y \mid X$] as well. This is also understandable since we have to consider both directions EXZ and EZX before we can decide which is the confounder (the "confusus directionis").

Conditional independence or relative freedom - continued

We took r = P[$S \mid C, \neg F$] as the basic risk and b = P[$S \mid \neg C, \neg F$] as the background risk. How is this related to independence ? When do the following equalities hold, and what would it mean ? Above statistical independence was not enough and we might be required to impose even stronger conditions.

$$r = \text{P}[S \mid C, \neg F] = ? = \text{P}[S \mid C] = R$$

$$b = \text{P}[S \mid \neg C, \neg F] = ? = \text{P}[S \mid \neg C] = B$$

It appears (see also the examples above) that we still have freedom to deviate from these equalities, even under independence. So, we can simply impose those conditions, as separate assumptions of their own.

Hence:

(i) Iff ($S \perp \neg F \mid C$) then $R = r$

(ii) Iff ($S \perp \neg F \mid \neg C$) then $B = b$

(iii) Iff ($S \perp \neg F \mid Truth$) then $R = r$, $B = b$, $R_F = p\, r + (1 - p)\, b$ and then the seeming relative risk is

$$\frac{R_F}{B_F} = \frac{p\, r + (1-p)\, b}{q\, r + (1-q)\, b}$$

(iv) If (iii) is extended to marginal independence of *Effect* and *Confounding* then this causes that the seeming relative risk must be 1:

$\frac{R_F}{B_F} = 1$

(v) Marginal independence of *Truth* and *Confounding* ($p = q = c$) of course is conceptually different from the notion of the relative freedom of S and $\neg F$ from *Truth*. (Different variables are involved.)

PM: Imposing relative freedom also has these consequences:

For (i), we already had $P[S, C, \neg F] = P[S \mid C, \neg F]\, P[C \mid \neg F]\, P[\neg F] = r\, q\, (1 - f)$. If it holds then $R = r$ and since $c = p\, f + q\, (1 - f)$:

$$P[S, C, F] + P[S, C, \neg F] = P[S, C] = R\, c = r\, c$$

$P[S \mid C, F] = P[S, C, F] / (P[C \mid F]\, P[F]) = (R\, c - r\, q\, (1 - f)) / (p\, f) = r\, (c - q\, (1 - f)) / (p\, f) = r$

For (ii), we had $P[S, \neg C, \neg F] = b\, (1 - q)\, (1 - f)$. If it holds then $B = b$ then similary $P[S \mid \neg C, F] = b$.

Considering the case when $R = r$ and $B = b$

That $R = r$ and $B = b$ is actually the situation studied by Schield (2003), "Simpson's paradox and Cornfield's conditions". In a nutshell, when translating epidemiology to terms and concepts that a simple economist like this author can understand, we had to develop the full apparatus above, to arrive at this special case. It hasn't been an easy path, and the relatively few pages for the reader above actually represent some years of study for this author. These apparantly are assumptions that epidemiologists may commonly make and that they might mention superficially but perhaps not too clearly for the cross-over scientist. The translation problem hinges on the point that in the "structural equations modelling" world, that forms the habitat of this author, conditional independences of course are used, but they are not discussed like we have done above, and the language and conventions of epidemiology didn't allow the quick connection as has been provided by the bridge above.

Relabel the parameters into the notation of Schield (2003):

$$\{r_1, r_2\} = \{r, b\} = \{R, B\}$$

$$\{R_1, R_2\} = \{R_F, B_F\}$$

$\{p_1, p_2\} = \{p, q\}$.

Then we can find the seeming risks as $R_i = \{r_1, r_2\}$. $\{p_i, 1 - p_i\}$, or $R_1 = p_1 r_1 + (1 - p_1)\, r_2$ and $R_2 = p_2\, r_1 + (1 - p_2)\, r_2$. The seeming risk difference is $R_1 - R_2 == (r_1 - r_2)(p_1 - p_2)$. If that difference is zero then the seeming relative risk is 1 and then effect and confounding are marginally independent.

A proportionality condition (since we assume constant rates) is that if $r_1 > r_2$ (it is a real cause, otherwise define the reverse) and $p_1 > p_2$ (it is a serious confounder, otherwise define the reverse) then also (3) $R_1 > R_2$ and, importantly, $p_1 / p_2 \geq R_1 / R_2$ (equal when $B = 0$).The latter inequality may be called the "Cornfield condition" (Schield (2003) and **Appendix B**). It may also be seen as $R - B \geq R_F - B_F$.

This sets $R = r$ and $B = b$.

```
lis = RiskToETCArray[{c, r, b}, {f, q}, {r, b}];
```

```
TableForm[lis, TableHeadings → CT["ETC", TableHeadings]] // Simplify
```

		Cause	¬ Cause
Success	Confounder	$(c + (f-1)\, q)\, r$	$b\,(-c + f - f\, q + q)$
	¬ Confounder	$-(f-1)\, q\, r$	$b\,(f-1)\,(q-1)$
¬ Success	Confounder	$-(c + (f-1)\, q)\,(r-1)$	$(b-1)\,(c + f\,(q-1) - q)$
	¬ Confounder	$(f-1)\, q\,(r-1)$	$-(b-1)\,(f-1)\,(q-1)$

```
(res = ETCStatistics[lis,  N → False] )  // MatrixForm
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	$q\, r$	$b - b\, q$	$-q\, b + b + q\, r$
¬ Success	$q - q\, r$	$(b-1)\,(q-1)$	$b\,(q-1) - q\, r + 1$
Sum	q	$1 - q$	1

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	$c\, r$	$b - b\, c$	$-c\, b + b + c\, r$
¬ Success	$c - c\, r$	$(b-1)\,(c-1)$	$b\,(c-1) - c\, r + 1$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	$c + (f-1)\, q$	$-c + f - f\, q + q$	f
¬ Confounder	$q - f\, q$	$(f-1)\,(q-1)$	$1 - f$
Sum	c	$1 - c$	1

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder
Success	$b\,(-c+f-f\,q+q)+(c+(f-1)\,q)\,r$	$(f-1)\,(b\,(q-1)-q\,r)$
¬ Success	$-q\,r\,f+f+b\,(c+f\,(q-1)-q)+(q-c)\,r$	$-(f-1)\,(b\,(q-1)-q\,r+1)$
Sum	f	$1-f$

$\left(\begin{array}{l}
N \to 1 \\
\text{NSuccess} \to -c\,b+b+c\,r \\
\text{NCause} \to c \\
\text{NConfounder} \to f \\
\text{IndependentPr(Truth, Confounding)} \to \frac{c-q}{f} = 0 \\
(\text{Success} \perp \neg\,\text{Confounder})(\text{Cause}) \to \text{True} \\
(\text{Success} \perp \neg\,\text{Confounder})(\neg\,\text{Cause}) \to \text{True} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause},\,\neg\,\text{Confounder}] \to r \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause},\,\neg\,\text{Confounder}] \to b \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause},\,\text{Confounder}] \to r \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause},\,\text{Confounder}] \to b \\
\text{Risk} \to \begin{pmatrix} r & b \\ r & b \end{pmatrix} \\
\text{Interaction} \to \{\text{Add} \to 0,\,\text{Times} \to 0\} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}] \to r \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause}] \to b \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\text{Confounder}] \to \frac{c+(f-1)\,q}{f} \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\neg\,\text{Confounder}] \to q \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Confounder}] \to \frac{b\,(-c+f-f\,q+q)+(c+(f-1)\,q)\,r}{f} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Confounder}] \to -q\,b+b+q\,r \\
\text{RRisk(True)} \to \frac{r}{b} \\
\text{RRisk(Cause)} \to \frac{r}{b} \\
\text{RelativePr(Confounder)} \to \frac{c+(f-1)\,q}{f\,q} \\
\text{RRisk(Seeming)} \to \frac{b\,(-c+f-f\,q+q)+(c+(f-1)\,q)\,r}{f\,(-q\,b+b+q\,r)} \\
\text{ETCAdjustedRRisk} \to \left\{\frac{r}{b},\,\frac{r}{b},\,\frac{r}{b},\,\frac{(1-f)\,r}{b}+\frac{f\,r}{b}\right\} \\
\text{Conditions} \to \left\{r>b,\,r>b,\,\frac{c-q}{f} \geq 0,\,\frac{(c-q)\,(b-r)}{f} \leq 0,\,\frac{b\,(q-c)}{f\,q\,(b\,(q-1)-q\,r)} \geq 0\right\} \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\text{Cause},\,\neg\,\text{Confounder}] \to 1-r \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\neg\,\text{Cause},\,\neg\,\text{Confounder}] \to 1-b \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\text{Cause},\,\text{Confounder}] \to 1-r \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\neg\,\text{Cause},\,\text{Confounder}] \to 1-b \\
\text{Safety} \to \begin{pmatrix} 1-r & 1-b \\ 1-r & 1-b \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} 1-r=0 & 1-b=1 \\ 1-r=0 & 1-b=1 \end{pmatrix} \\
\text{ETCSimpson} \to \{\text{Necessary} \to \text{False},\,\text{Sufficient} \to \{r>b,\,r>b,\,b>r\}\}
\end{array}\right)$

The difference now also is in the middle column.

```
ETCRiskTable[res] // Rationalize // Simplify
```

	Name	Value	Name	Value	Name	Value
Cause	r	r	R	r	Rf	$\frac{b(-c+f-f q+q)+(c+(f-1)q)r}{f}$
Background	b	b	B	b	Bf	$-q b+b+q r$
Difference	r – b	$r-b$	R – B	$r-b$	Rf – Bf	$-\frac{(c-q)(b-r)}{f}$
Ratio	r / b	$\frac{r}{b}$	R / B	$\frac{r}{b}$	Rf / Bf	$\frac{b(-c+f-f q+q)+(c+(f-1)q)r}{f(-q b+b+q r)}$

Though the above does not quite show it, the seeming relative risk R_F / B_F simplifies. To show this, we need to do the following.

- Note that the output above does not have an simple *p*.

  ```
  p == ConditionalPr["Cause"]["Confounder"] /. res
  ```

  $$p=\frac{c+(f-1)q}{f}$$

- But if we use it ... then we find the Cornfield et al. condition mentioned by Schield (2003).

  ```
  RRisk["Seeming"] /. res /. c → p f + q (1 – f) // Simplify
  ```

  $$\frac{-p b+b+p r}{-q b+b+q r}$$

And by consequence for R_F:

```
ConditionalPr["Success"][ "Confounder"] /. res /. c → p f + q (1 – f) // Simplify
```

$$-p b+b+p r$$

Technical note: If *p* were introduced in the input then the input would be overdetermined. An option might be to let *n* become the dependent outcome but then we would not have a normalized situation (or have all kinds of checks and possibly arbitrary internal solutions on it). The current input format seems optimal, with the small cost that the seeming relative risk looks a bit differently than the ratio of weighted rates. It is a bit unfortunate that current mathematical concepts and routines are awkward at handling so-called “overdetermined” situations that however are neat human psychological ways to handle information. (This would be another opportunity to devise a notation to capture this notion.)

Variations on input

Consider the example contingency table that this discussion started with. Given the output from the statistical analysis we can easily re-create the figures in the table with our current method at parameterization. Let us consider some variations.

The average relative risk R / B for the disease is about 9. For the confounder we keep the 800 versus 200 split, so that $f = 0.8$. Of the confounding group some $p = 51\%$ of their numbers contribute to the risk population and the non-confounders contribute $q = 24\%$ of their numbers. The prevalence c of the disease then becomes:

```
c == 0.2 * .24 + 0.8 * .51
```

$c = 0.456$

The ETCStatistics already created key output. We can substitute these in the right slots. We put the routine in Hold, otherwise we would just recover the same data (except that 1000.0 should be an integer and not a real).

```
RiskToETCArray[] /. res1
```

Hold[RiskToETCArray][{0.456, 0.145833, 0.0328947}, {0.8, 0.24}, {0.179825, 0.0202206}, 1000.]

We already observed above that the example contingency table does not satisfy the assumptions of conditional independence since we find that $r \neq R$ *and* $b \neq B$. What is an interesting variation on the input ? Normally we would take the r and b since these would be the key parameters. But this is an exercise, and thus we may also take R and B and see what happens.

- This is what happens when we take R and B, that have a seeming relative risk $R / B = 9$. Something goes horribly wrong.

```
lis = RiskToETCArray[{.456, .18, .02}, {0.8, .24}, {.18, 02}, 1000];
```

ETCArrayCheck::neg : Negative elements found R < B

ETCArrayCheck::rel : Warning: p / q < Rf / Bf

ETCArrayCheck::neg : Negative elements found {−693}

- And it does not help if we scale it down.

```
lis = RiskToETCArray[{.456, .09, .01}, {0.8, .24}, {.09, 01}, 1000];
```

ETCArrayCheck::neg : Negative elements found R < B

ETCArrayCheck::rel : Warning: p / q < Rf / Bf

ETCArrayCheck::neg : Negative elements found {−150}

- So, let us just take the structural parameters (as we already planned to do).

```
lis = RiskToETCArray[{.456, .1458, .0329}, {0.8, .24}, {.1458, .0329}, 1000];
```

```
TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	59	13
	¬ Confounder	7	5
¬ Success	Confounder	349	379
	¬ Confounder	41	147

The current table is based upon conditional independence while the original example wasn't. It is hard to say what the differences amount to - especially since these are only arbitrary numbers.

- These are the differences with respect to the example table. The results are no different when the confounder isn't present. When it is present, we lose some successes, most when the cause is present, with some compensation when the cause is't present.

```
TableForm[lis − CT[Data], TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	−16	7
	¬ Confounder	0	0
¬ Success	Confounder	16	−7
	¬ Confounder	0	0

The summary statistics become:

- There is a small deviation from conditional independence since we rounded the contingency table. A useful point to note is that relative freedom is not sufficient to turn this case into one of "simple causality". The "cause" that we have here is only a "contributing factor".

```
(Report[Variant] = ETCStatistics[lis, N → True] // N) // MatrixForm
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	0.035	0.025	0.06
¬ Success	0.205	0.735	0.94
Sum	0.24	0.76	1.

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	66	18	84
¬ Success	390	526	916
Sum	456	544	1000

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	408	392	800
¬ Confounder	48	152	200
Sum	456	544	1000

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder	Total
Success	72	12	84
¬ Success	728	188	916
Sum	800	200	1000

$$\left(\begin{array}{l}
N \to 1000. \\
\text{NSuccess} \to 84. \\
\text{NCause} \to 456. \\
\text{NConfounder} \to 800. \\
\text{MarginalPr(Success)} \to 0.084 \\
\text{MarginalPr(Cause)} \to 0.456 \\
\text{MarginalPr(Confounder)} \to 0.8 \\
\text{IndependentPr(Truth, Confounding)} \to \text{False} \\
(\text{Success} \perp \neg\, \text{Confounder})(\text{Cause}) \to \text{False} \\
(\text{Success} \perp \neg\, \text{Confounder})(\neg\, \text{Cause}) \to \text{False} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}, \neg\, \text{Confounder}] \to 0.145833 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\, \text{Cause}, \neg\, \text{Confounder}] \to 0.0328947 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}, \text{Confounder}] \to 0.144608 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\, \text{Cause}, \text{Confounder}] \to 0.0331633 \\
\text{Risk} \to \begin{pmatrix} 0.144608 & 0.0331633 \\ 0.145833 & 0.0328947 \end{pmatrix} \\
\text{Interaction} \to \{\text{Add} \to -0.00149402, \text{Times} \to -0.0728507\} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}] \to 0.144737 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\, \text{Cause}] \to 0.0330882 \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\text{Confounder}] \to 0.51 \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\neg\, \text{Confounder}] \to 0.24 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Confounder}] \to 0.09 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\, \text{Confounder}] \to 0.06 \\
\text{RRisk(True)} \to 4.43333 \\
\text{RRisk(Cause)} \to 4.37427 \\
\text{RelativePr(Confounder)} \to 2.125 \\
\text{RRisk(Seeming)} \to 1.5 \\
\text{ETCAdjustedRRisk} \to \{4.37427, 4.43333, 4.36048, 4.37505\} \\
\text{Conditions} \to \{\text{True}, \text{True}, \text{True}, \text{True}, \text{True}\} \\
\text{ConditionalPr}[\,\neg\, \text{Success}\,][\,\text{Cause}, \neg\, \text{Confounder}] \to 0.854167 \\
\text{ConditionalPr}[\,\neg\, \text{Success}\,][\,\neg\, \text{Cause}, \neg\, \text{Confounder}] \to 0.967105 \\
\text{ConditionalPr}[\,\neg\, \text{Success}\,][\,\text{Cause}, \text{Confounder}] \to 0.855392 \\
\text{ConditionalPr}[\,\neg\, \text{Success}\,][\,\neg\, \text{Cause}, \text{Confounder}] \to 0.966837 \\
\text{Safety} \to \begin{pmatrix} 0.855392 & 0.966837 \\ 0.854167 & 0.967105 \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} \text{False} & \text{False} \\ \text{False} & \text{False} \end{pmatrix} \\
\text{ETCSimpson} \to \{\text{Necessary} \to \text{False}, \text{Sufficient} \to \{\text{True}, \text{True}, \text{False}\}\}
\end{array}\right)$$

The main conclusion is that the average relative risk is no longer 9. NB. We should have $R / B = r / b$ but after constructing the data matrix we rounded the data again so there is a small difference.

ETCRiskTable[]

	Name	Value	Name	Value	Name	Value
Cause	r	0.145833	R	0.144737	Rf	0.09
Background	b	0.0328947	B	0.0330882	Bf	0.06
Difference	r – b	0.112939	R – B	0.111649	Rf – Bf	0.03
Ratio	r / b	4.43333	R / B	4.37427	Rf / Bf	1.5

Thus, imposing relative freedom makes that the average relative risk becomes equal to the true relative risk, $R / B = r / b$. Either (1) we get negative values (if we impose an average relative risk of 9) or (2) we accept the true relative risk but then see the average relative risk adjusted.

OutsideTable[Report, {Example, Variant},
{RRisk["True"], RRisk["Cause"], RelativePr["Confounder"], RRisk["Seeming"]}]

	Example	Variant
RRisk(True)	4.43333	4.43333
RRisk(Cause)	8.89314	4.37427
RelativePr(Confounder)	2.125	2.125
RRisk(Seeming)	1.6875	1.5

Variation would rather be done on the true parameters and not on the averages that are found. To take the averages or their ratio's as the true parameters for this 2 × 2 × 2 table might be called the "confusus additionis" (a special case of "confusus magnitudinis").

Switching between truth and confounding

We mentioned the issue of choosing EXZ or EZX. We may read "cause" as "confounder" and conversely, and repeat the analysis. One hopes that this does not confound the reader.

- This still uses the labels as above. It puts the confounder in the middle of the table so that a call of the ETCStatistics routine will take it as the cause.

CT[Order, {"Effect", "Confounding", "Truth"}]

		Confounder	¬ Confounder
Success	Cause	75	7
	¬ Cause	6	5
¬ Success	Cause	333	41
	¬ Cause	386	147

If we take the statistics of this case it will appear that $r < b$ *or* $R < B$. This makes for a dumb causal model and silly confounding. Hence, we reverse the categories as well. When good weather has a lower risk than bad weather, then the latter should be the true

cause. We can try various reversals for various variables untill we have a serious causal model.

- This is from above model.

```
(Report[Confound] = ETCStatistics[%, N → True, Print → False] // N) //
    MatrixForm;
```

```
"Conditions" /. Report[Confound]
```

{False, True, True, True, False}

- This reverses the old confounder and new cause categories. Still some conditions not satisfied.

```
CT[Switch, "NewModel−1", "ETC−1−Truth−1−Confounding−1−Effect",
    "Confounding" → {! "Confounder", "Confounder"}];
```

CT::cop : Label NewModel−1 already known in CT[*List*]

```
(Report[Confound] = ETCStatistics[CT["NewModel−1", Data],
          N → True, Print → False] // N) // MatrixForm;
```

```
"Conditions" /. Report[Confound]
```

{True, False, False, True, False}

- This reverses also the supposed confounder categories. Still some conditions not satisfied.

```
CT[Switch, "NewModel−2",
    "ETC−1−Truth−1−Confounding−1−Effect", "Truth" → {! "Cause", "Cause"},
    "Confounding" → {! "Confounder", "Confounder"}];
```

CT::cop : Label NewModel−2 already known in CT[*List*]

```
(Report[Confound] = ETCStatistics[CT["NewModel−2", Data],
          N → True, Print → False] // N) // MatrixForm;
```

```
"Conditions" /. Report[Confound]
```

{False, False, True, False, True}

- This reverses the categories for all variables, also the effect. Ah, finally all conditions are satisfied.

```
CT[Switch, "NewModel−3", "ETC−1−Truth−1−Confounding−1−Effect",
  "Effect" → {! "Success", "Success"}, "Truth" → {! "Cause", "Cause"},
  "Confounding" → {! "Confounder", "Confounder"}]
```

CT::cop : Label NewModel−3 already known in CT[*List*]

```
(Report[Confound] = ETCStatistics[CT["NewModel−3", Data],
          N → True, Print → False] // N) // MatrixForm;
```

```
"Conditions" /. Report[Confound]
```

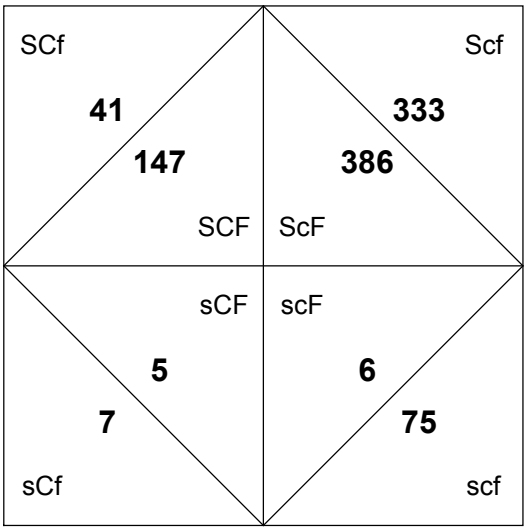
{True, True, True, True, True}

We have found a new model that on the face of it might be a causal model while we also might be properly confounded. Above table still contains the original labels. We now relabel. Above runs suppressed the output but we can show it now.

With these relabelled data, the analysis would be that the "the absence of the original confounder" would cause the non-success. An example helps. If the original model EXZ for example would be that smoking caused lung cancer, with a confounding difference between cities (bad air) and rural areas (good air), then the new causal analysis EZX would be that the rural areas with their good air "caused" the absence of lung cancer. We are not speaking about statistical association but about cause here. Let us first produce the statistics and then think about the causal chain (as people normally tend to do).

- Now everything is relabeled. What was *Truth* now is *Confouding*, what was ¬*F* now is *C,* and so on.

```
ETCSquare["NewModel−3"];
```



- This also uses the new labels.

```
(Report[Confound] = ETCStatistics[CT["NewModel−3", Data], N → True] // N) //
  MatrixForm
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	0.0899123	0.730263	0.820175
¬ Success	0.0153509	0.164474	0.179825
Sum	0.105263	0.894737	1.

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	188	719	907
¬ Success	12	81	93
Sum	200	800	1000

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	152	392	544
¬ Confounder	48	408	456
Sum	200	800	1000

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder	Total
Success	533	374	907
¬ Success	11	82	93
Sum	544	456	1000

$$\left(\begin{array}{l}
N \to 1000. \\
\text{NSuccess} \to 907. \\
\text{NCause} \to 200. \\
\text{NConfounder} \to 544. \\
\text{MarginalPr(Success)} \to 0.907 \\
\text{MarginalPr(Cause)} \to 0.2 \\
\text{MarginalPr(Confounder)} \to 0.544 \\
\text{IndependentPr(Truth, Confounding)} \to \text{False} \\
(\text{Success} \perp \neg\ \text{Confounder})(\text{Cause}) \to \text{False} \\
(\text{Success} \perp \neg\ \text{Confounder})(\neg\ \text{Cause}) \to \text{False} \\
\text{ConditionalPr}[\text{ Success }][\text{ Cause, } \neg\ \text{Confounder}] \to 0.854167 \\
\text{ConditionalPr}[\text{ Success }][\neg\ \text{Cause, } \neg\ \text{Confounder}] \to 0.816176 \\
\text{ConditionalPr}[\text{ Success }][\text{ Cause, Confounder}] \to 0.967105 \\
\text{ConditionalPr}[\text{ Success }][\neg\ \text{Cause, Confounder}] \to 0.984694 \\
\text{Risk} \to \begin{pmatrix} 0.967105 & 0.984694 \\ 0.854167 & 0.816176 \end{pmatrix} \\
\text{Interaction} \to \{\text{Add} \to -0.0555788, \text{Times} \to -0.0644086\} \\
\text{ConditionalPr}[\text{ Success }][\text{ Cause}] \to 0.94 \\
\text{ConditionalPr}[\text{ Success }][\neg\ \text{Cause}] \to 0.89875 \\
\text{ConditionalPr}[\text{ Cause }][\text{ Confounder}] \to 0.279412 \\
\text{ConditionalPr}[\text{ Cause }][\neg\ \text{Confounder}] \to 0.105263 \\
\text{ConditionalPr}[\text{ Success }][\text{ Confounder}] \to 0.979779 \\
\text{ConditionalPr}[\text{ Success }][\neg\ \text{Confounder}] \to 0.820175 \\
\text{RRisk(True)} \to 1.04655 \\
\text{RRisk(Cause)} \to 1.0459 \\
\text{RelativePr(Confounder)} \to 2.65441 \\
\text{RRisk(Seeming)} \to 1.1946 \\
\text{ETCAdjustedRRisk} \to \{1.0459, 1.04655, 0.982138, 1.01151\} \\
\text{Conditions} \to \{\text{True, True, True, True, True}\} \\
\text{ConditionalPr}[\neg\ \text{Success }][\text{ Cause, } \neg\ \text{Confounder}] \to 0.145833 \\
\text{ConditionalPr}[\neg\ \text{Success }][\neg\ \text{Cause, } \neg\ \text{Confounder}] \to 0.183824 \\
\text{ConditionalPr}[\neg\ \text{Success }][\text{ Cause, Confounder}] \to 0.0328947 \\
\text{ConditionalPr}[\neg\ \text{Success }][\neg\ \text{Cause, Confounder}] \to 0.0153061 \\
\text{Safety} \to \begin{pmatrix} 0.0328947 & 0.0153061 \\ 0.145833 & 0.183824 \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} \text{False} & \text{False} \\ \text{False} & \text{False} \end{pmatrix} \\
\text{ETCSimpson} \to \{\text{Necessary} \to \text{False, Sufficient} \to \{\text{False, True, True}\}\}
\end{array}\right)$$

ETCRiskTable[]

	Name	Value	Name	Value	Name	Value
Cause	r	0.854167	R	0.94	Rf	0.979779
Background	b	0.816176	B	0.89875	Bf	0.820175
Difference	r − b	0.0379902	R − B	0.04125	Rf − Bf	0.159604
Ratio	r / b	1.04655	R / B	1.0459	Rf / Bf	1.1946

This reproduces the output of the original example EXZ and the new EZX.

OutsideTable[Report, {Example, Confound}, {RRisk["True"], RRisk["Cause"], RelativePr["Confounder"], RRisk["Seeming"], "Conditions"}]

	Example	Confound
RRisk(True)	4.43333	1.04655
RRisk(Cause)	8.89314	1.0459
RelativePr(Confounder)	2.125	2.65441
RRisk(Seeming)	1.6875	1.1946
Conditions	True True True True True	True True True True True

The numerical conditions for calling this a proper causal model are satisfied. Yet, in terms of content it may not quite convince. Of course, it all depends on what the content is. As in the stated example of lung cancer and rural areas, it is difficult to imagine how people would have a natural disposition to cancer and that "good rural air" would prevent or cure it. It might be some curative pollen or so. In terms of causality, it is turning the world upside down in changing the causality that smoking and bad air cause lung cancer into that rural good air causes health (with non-smoking as a confouder). But, of course, these are entirely fictional data and we don't have a real problem so we cannot say anything yet. The only conclusion that we arrive at is that these conditions and manipulations definitely can help and guide us towards better understanding the causal relations. For a definite answer on causality we still depend upon the true model of the world.

Safety and the conditions on safety are worth mentioning too. The differences from 0 (left column) and 1 (right column) make that we do not have a simple cause in this EZX configuration. The "wunderbar" result of the original has been greatly resolved since w is much closer to 0 now. But it comes at the price of a v that was close to 1 as it should be and now is closer to 0 as it shouldn't be. The lack of evidence for a EZX would support the EXZ interpretation.

Some readers might expect that I now reveal that the true relation was EZX to start with. Sorry, these are entirely fictional data.

Comparing the Simpson paradox and the Cornfield et al. condition

The Cornfield et al. condition arises from assuming relative freedom or conditional independence of S and F given C. Thus, a variable is defined to be a "confounder" if and only if it cannot contribute information on the causal relation from C to S. This is a rather strong condition since it limits the degree to which we can be confounded. This could be called the "confusus confusi". However, when the condition is satisfied, which we can do by directly checking conditional independence or using the inequality condition derived by Cornfield et al. (easiest $R - B \geq R_F - B_F$), then we may indeed have more confidence in the notion that F is only a confounder. However, we would still need a model to explain the causal relations, since we cannot exclude mere chance as the reason that the condition is satisfied.

The Simpson paradox as discussed in **Appendix A** is a bit different. The main thrust of the paradox comes from subpopulations such that those show property A while the summed total shows property $\neg A$. The particularly relevant property is relative risk, such that the subpopulations have RR > 1 while RR < 1 for the summed total. The only two serious subpopulations that we have in the ETC world comes from the division in F and $\neg F$. If we were to divide along the lines of Effect and Truth then this would not make sense since it is precisely their relationship that is the causal one. What happens with their properties and proportions is not relevant and thus cannot be paradoxical. Taking the division along the line of Confounding makes sense in that the distribution of f and $1 - f$ is taken as a more or less "causal" explanation for the overall effect. Thus, the only relevant Simpson paradox for the ETC world is that the subpopulations of F and $\neg F$ have relative risks pointing one way while the total points another way. The final question is what directions to take. For the $\neg F$ population we have already assumed that $r > b$, so that RR > 1. It would be strange for the F population to have a different direction, and thus we have RR > 1 there too. Hence, the paradox would be that the total would show RR < 1. That would be a paradox indeed, since under our assumptions the confounder really cannot affect the true causal relations so that it would be surprising indeed if it were to affect the relative risk measure. In **Appendix A** we derive for the paradox to occur: (i) the necessary conditions for the paradox as $b < r < 1 - v < 1 - w$, which translates too as $w < v < 1 - r$, and (ii) the sufficient condition that when these are satisfied and then $R < B$. In the discussion above on safety we determined the ranges of the parameters. For a causal process to get closer to the simple causal model we would

require that $w \to 0$, $v \to 1$ and $r \to 1$. The causal model requires that $v \to 1$ while the Simpson paradox requires that $v \to 0$. Under normal causal assumptions that $v > 1 - r$ the Simpson paradox could not exist. The Simpson paradox requires that the "other causes" would have perverse effects. The outcome $R < B$ makes less sense when we are speaking about a true cause and a true confounder. Hence, in the $2 \times 2 \times 2$ ETC world the Simpson paradox has no good reason to exist, and if the conditions are satisfied then one should check one's causal model. What is conceivable, however, is a relative effect, in that the subpopulations have say RR > 3 and the sum 1 < RR < 3. This has not been looked into.

The Cornfield condition is sufficient to prevent Simpson too. Under relative freedom or conditional independence $\{R, B\} = \{r, b\}$ and with $r > b$ we find that the Simpson paradox cannot occur. Yet this limits the range of possible models. It is more adequate to allow for the possibility of (some, statistical) dependence. It makes more sense to require that $v > 1 - r$ and this indeed also prevents the paradox.

We might call it the "confusus conditionis (Simpson et Cornfield et al.)" to not see these conditions in their proper relation.

If some crucial data are missing

Up to now we have been assuming that we have a completely filled contingency table. It may also be that the crucial second line is missing so that we cannot calculate our r and b. As said, experimental economics may meet with practical or moral limitations. For example, if we do an experiment on the impact of drinking water or drinking beer on the quality of decisions by Central Bankers, all other drinks excluded, then it might not be considered appropriate to withhold them those two drinks as well for the weeks that the experiment would take.

```
lis = CT[Data] /. {5 → Missing[1], 147 → Missing[2]};

TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	75	6
	¬ Confounder	7	Missing(1)
¬ Success	Confounder	333	386
	¬ Confounder	41	Missing(2)

There are all kinds of variations on this theme. For example, the data on success may come from one country and the data from failure may come from another country, and one has to patch together a joint story. Or, indeed, we would leave the $2 \times 2 \times 2$ world

and meet more variables and sizes. All this is just to say that the ETC model looks strong but that this may be deceptive since it assumes full knowledge.

We cannot review all possible combinations on missing data but now that we have mentioned above example it is tempting to consider it. Let us compare the data table with the fully parameterized table. However, the most crucial missing datum is the total sample size *n.* For a controlled experiment one might set the total number of cases where both cause and confounder are withheld, so the *m* = Missing[1] + Missing[2] is known but not how it is distributed over success or failure. In that case, also, *m* is set by the experimenter to an arbitrary number, and then not actually performed, which has a ring of magic to it, since we may choose any number and then not do it. When the experiment however runs over a limited number of weeks then one can indeed imagine that *m* has a proper value. Let us assume this, and replace Missing[2] such that the total number is 1000 again.

```
lis2 = SafetyToETCArray[{c, r, b}, {f, q}, {w, v}] 1000;
```

```
TableForm[lis2, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	$1000\,(c-(1-f)\,q)\,(1-w)$	$1000\,(-c+f\,(1-q)+q)\,(1-v)$
	¬ Confounder	$1000\,(1-f)\,q\,r$	$1000\,b\,(1-f)\,(1-q)$
¬ Success	Confounder	$1000\,(c-(1-f)\,q)\,w$	$1000\,(-c+f\,(1-q)+q)\,v$
	¬ Confounder	$1000\,(1-f)\,q\,(1-r)$	$1000\,(1-b)\,(1-f)\,(1-q)$

```
sol = Solve[Add[lis] == 1000, Missing[2]];
```

```
lis = CT[Data] /. {5 → Missing[1], 147 → Missing[2]} /. sol[[1]];
```

```
TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	75	6
	¬ Confounder	7	Missing(1)
¬ Success	Confounder	333	386
	¬ Confounder	41	152 – Missing(1)

It appears that other key parameters are not affected and that mainly our estimate of *b* is impossible. The data still allow the calculation of *c, f* and *q,* still unaffected at 0.456, 0.8 and 0.24, so that the Missing[1] value is direct in proportion to *b,* as Missing[1] = 1000 *b* * 0.2 * 0.76. Thus, in the statistics also those ratios are affected that depend upon *b.*

```
(res = (ETCStatistics[lis] // N) /. (1. → 1)) // MatrixForm
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	0.035	0.005 Missing(1.)	0.005 (Missing(1.) + 7.)
¬ Success	0.205	0.005 (152. − 1. Missing(1.))	0.005 (193. − 1. Missing(1.))
Sum	0.24	0.76	1.

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	82	Missing(1) + 6	Missing(1) + 88
¬ Success	374	538 − Missing(1)	912 − Missing(1)
Sum	456	544	1000

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	408	392	800
¬ Confounder	48	152	200
Sum	456	544	1000

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder	Total
Success	81	Missing(1) + 7	Missing(1) + 88
¬ Success	719	193 − Missing(1)	912 − Missing(1)
Sum	800	200	1000

$$\left\{\begin{array}{l}
N \to 1000. \\
\text{NSuccess} \to \text{Missing}(1) + 88. \\
\text{NCause} \to 456. \\
\text{NConfounder} \to 800. \\
\text{MarginalPr}(\text{Success}) \to 0.001\ (\text{Missing}(1) + 88.) \\
\text{MarginalPr}(\text{Cause}) \to 0.456 \\
\text{MarginalPr}(\text{Confounder}) \to 0.8 \\
\text{IndependentPr}(\text{Truth}, \text{Confounding}) \to \text{False} \\
(\text{Success} \perp \neg\ \text{Confounder})(\text{Cause}) \to \text{False} \\
(\text{Success} \perp \neg\ \text{Confounder})(\neg\ \text{Cause}) \to 49.\ \text{Missing}(1) = 114. \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \text{Cause}, \neg\ \text{Confounder}] \to 0.145833 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \neg\ \text{Cause}, \neg\ \text{Confounder}] \to 0.00657895\ \text{Missing}(1) \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \text{Cause}, \text{Confounder}] \to 0.183824 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \neg\ \text{Cause}, \text{Confounder}] \to 0.0153061 \\
\text{Risk} \to \begin{pmatrix} 0.183824 & 0.0153061 \\ 0.145833 & 0.00657895\ \text{Missing}(1) \end{pmatrix} \\
\text{Interaction} \to \left\{\text{Add} \to 0.00657895\ \text{Missing}(1) + 0.0226841,\ \text{Times} \to 12.0098 - \frac{22.1667}{\text{Missing}(1)}\right\} \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \text{Cause}] \to 0.179825 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \neg\ \text{Cause}] \to 0.00183824\ (\text{Missing}(1) + 6.) \\
\text{ConditionalPr}[\ \text{Cause}\ ][\ \text{Confounder}] \to 0.51 \\
\text{ConditionalPr}[\ \text{Cause}\ ][\ \neg\ \text{Confounder}] \to 0.24 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \text{Confounder}] \to 0.10125 \\
\text{ConditionalPr}[\ \text{Success}\ ][\ \neg\ \text{Confounder}] \to 0.005\ (\text{Missing}(1) + 7.) \\
\text{RRisk}(\text{True}) \to \frac{22.1667}{\text{Missing}(1)} \\
\text{RRisk}(\text{Cause}) \to \frac{97.8246}{\text{Missing}(1)+6.} \\
\text{RelativePr}(\text{Confounder}) \to 2.125 \\
\text{RRisk}(\text{Seeming}) \to \frac{20.25}{\text{Missing}(1)+7.} \\
\text{ETCAdjustedRRisk} \to \left\{\frac{97.8246}{\text{Missing}(1)+6.},\ \frac{22.1667}{\text{Missing}(1)},\ 12.0098,\ 9.60784 + \frac{4.43333}{\text{Missing}(1)}\right\} \\
\text{Conditions} \to \left\{6.\ \text{Missing}(1) < 133.,\ 57.\ \text{Missing}(1) < 5234.,\ \text{True},\ 4.\ \text{Missing}(1) \le 53.,\ \frac{20.25}{\text{Missing}(1)+7.}\right. \\
\text{ConditionalPr}[\ \neg\ \text{Success}\ ][\ \text{Cause}, \neg\ \text{Confounder}] \to 0.854167 \\
\text{ConditionalPr}[\ \neg\ \text{Success}\ ][\ \neg\ \text{Cause}, \neg\ \text{Confounder}] \to 1 - 0.00657895\ \text{Missing}(1) \\
\text{ConditionalPr}[\ \neg\ \text{Success}\ ][\ \text{Cause}, \text{Confounder}] \to 0.816176 \\
\text{ConditionalPr}[\ \neg\ \text{Success}\ ][\ \neg\ \text{Cause}, \text{Confounder}] \to 0.984694 \\
\text{Safety} \to \begin{pmatrix} 0.816176 & 0.984694 \\ 0.854167 & 1 - 0.00657895\ \text{Missing}(1) \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} \text{False} & \text{False} \\ \text{False} & 1 - 0.00657895\ \text{Missing}(1) = 1 \end{pmatrix} \\
\text{ETCSimpson} \to \{\text{Necessary} \to \text{False},\ \text{Sufficient} \to \{\text{True},\ 6.\ \text{Missing}(1) < 133.,\ \text{Missing}(1) < -766
\end{array}\right.$$

We just considered one consequence of missing data. The general idea is that when data are missing then this reduces the scope for conclusions. Perhaps there are more possibilities for conjectures like “if the process would be conditionally independent then

..."), but when such conjectures are not testable due to the lack of data anyhow, then there seems little value in them.

Testing the parameters

Above discussion allowed us to identify key parameters that arise when the cause is truly the cause and the confounder truly the confounder, so that we find the impact of the cause when the confounder is absent. Our analysis was based upon one "set of data" only. Presumably, contingency tables are created with the observations of data records $\{y, x, z\}$, and each record adds to a cell somewhere in the table. A set of data is created by a stopping rule, e.g. when we reach the required n. We might repeat the sampling and stop at some m. If our assumptions on the parameterization are correct then both crosstables would have the same structural parameters r, b, w and v and the variation would concentrate on c, f and possibly q. The assumption of constant w and v could be too quick though. Our assumption is that these parameters are more related to the cause than to the confounder, but they are established in the context when the confounder is present, so we may just beg the question. Let us consider varying those too. Thus, let us consider two cases, one where r, b, w and v are constant and one where only r and b are constant.

(1) Assuming that r, b, w, v are constant.

```
lis = SafetyToETCArray[{c1, r, b}, {f1, q1}, {w, v}] n;

lis2 = SafetyToETCArray[{c2, r, b}, {f2, q2}, {w, v}] m;

lis3 = lis + lis2;

TableForm[lis3, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	
Success	Confounder	$n(1-w)(c_1-(1-f_1)q_1)+m(1-w)(c_2-(1-f_2)q_2)$	$n(1-$
	¬ Confounder	$n\,r(1-f_1)q_1+m\,r(1-f_2)q_2$	$b\,n(1$
¬ Success	Confounder	$n\,w(c_1-(1-f_1)q_1)+m\,w(c_2-(1-f_2)q_2)$	
	¬ Confounder	$n(1-r)(1-f_1)q_1+m(1-r)(1-f_2)q_2$	

```
res = ETCStatistics[lis3, Print → False];
```

These points must be noted:

- We find that *q* is a weighed average of the two tables, with the absence frequencies as weights. If *q* is constant then it remains constant. If there is variation then there is a tendency to the mean.

```
ConditionalPr["Cause"][! "Confounder"] /. res
```

$$\frac{n\,(f_1-1)\,q_1+m\,(f_2-1)\,q_2}{f_2\,m-m-n+n\,f_1}$$

- For *w*:

```
ConditionalPr[! "Success"]["Cause", "Confounder"] /. res
```

w

- For *v*:

```
ConditionalPr[! "Success"][! "Cause", "Confounder"] /. res
```

v

This only holds when there is a "basic process" with parameters *r, b, w* and *v,* such that all observations are mere scaling-ups with the *c, f* and *q*.

(2) Above assumption that *w* and *v* are constant may be too strong. Assume that only *r* and *b* are constant.

```
lis = SafetyToETCArray[{c1, r, b}, {f1, q1}, {w1, v1}] n;

lis2 = SafetyToETCArray[{c2, r, b}, {f2, q2}, {w2, v2}] m;

lis3 = lis + lis2;

res = ETCStatistics[lis3, Print → False];
```

- The former conclusion of *q* is unaffected. Its value is not affected by the other parameters.

```
ConditionalPr["Cause"][! "Confounder"] /. res
```

$$\frac{n\,(f_1-1)\,q_1+m\,(f_2-1)\,q_2}{f_2\,m-m-n+n\,f_1}$$

- For *w,* we however find that it is a weighed average now. This remains so also when *q* is constant. The average outcome is affected by both *c* and *f,* and the mediator is *q,* which is not surprising since it affects the size of the subpopulation.

  ConditionalPr[! "Success"]["Cause", "Confounder"] /. res

  $$\frac{n\,c_1\,w_1 + n\,(f_1 - 1)\,q_1\,w_1 + m\,(c_2 + (f_2 - 1)\,q_2)\,w_2}{n\,c_1 + m\,c_2 - n\,q_1 + n\,f_1\,q_1 - m\,q_2 + m\,f_2\,q_2}$$

  % /. { q_1 → q, q_2 → q} // Simplify

  $$\frac{n\,(c_1 + q\,(f_1 - 1))\,w_1 + m\,(c_2 + q\,(f_2 - 1))\,w_2}{n\,c_1 + m\,c_2 + q\,(f_2\,m - m - n + n\,f_1)}$$

  % /. {c_1 → $p_1\,f_1 + q_1\,(1 - f_1)$, c_2 → $2\,f_2 + 2\,(1 - f_2)$} // Simplify

  $$\frac{n\,(-q + f_1\,(q + p_1 - q_1) + q_1)\,w_1 + m\,(f_2\,q - q + 2)\,w_2}{-q\,m + q\,f_2\,m + 2\,m - n\,q + n\,f_1\,(q + p_1 - q_1) + n\,q_1}$$

- For *v,* the same conclusion as for *w.*

  ConditionalPr[! "Success"][! "Cause", "Confounder"] /. res

  $$\frac{n\,c_1\,v_1 + n\,f_1\,(q_1 - 1)\,v_1 - n\,q_1\,v_1 + m\,c_2\,v_2 - m\,f_2\,v_2 - m\,q_2\,v_2 + m\,f_2\,q_2\,v_2}{n\,c_1 + m\,c_2 - n\,f_1 - m\,f_2 - n\,q_1 + n\,f_1\,q_1 - m\,q_2 + m\,f_2\,q_2}$$

  % /. { q_1 → q, q_2 → q} // Simplify

  $$\frac{n\,(-q + c_1 + (q - 1)\,f_1)\,v_1 + m\,(-q + c_2 + (q - 1)\,f_2)\,v_2}{-m\,q - n\,q + n\,f_1\,q + m\,f_2\,q + n\,c_1 + m\,c_2 - n\,f_1 - m\,f_2}$$

By conclusion, we have the typical estimation problem that two data subsets give different estimates and that the joint set gives some average. There is nothing particularly worrying about it, and generally we would hold that the larger data set gives the best estimate. The breakdown does allow us however to test the constancy of *q,* and its consequences for the other parameters.

After considering these two cases, two final points to note are:

(i) When we consider a second confounder F_2 then we would find the true impact parameters from absence of it. Thus, optimally, above discussion on the ETC table is under the assumption that a second confounder is absent.

(ii) Our problem in estimation might not be to recover the basic risks but to give an "overall outcome" given that the confounder is present. Let us link up to the issue of relative risk, both crude and adjusted. Once we have identified the crucial parameters {*r,*

b, w, v} we basically have caught the causal process, except for the prevalences *c* and *f,* and of course the statistical link *q* between them. Would we have any need for an "(adjusted) relative risk" measure ? In all likelihood we would, for important psychological reasons. As we come from a 2 × 2 world we take the *R / B* as our frame of reference but we discover that the true relative risk is *r / b* (controlling for the confounder, i.e. looking when it is absent). We may easily adapt our perception of the true risk involved, yet, we also are stuck with the possibility that the confounder might be present, in which case the relative risk gets the value RR_F = (1 - *w*) / (1 - *v*). It remains to be seen whether *R / B* or the artificial AdjRR = (1 - *f*) *r / b* + *f* RR_F gives the accurate description of the "overall risk". But it must be admitted that when we have been trained to think in terms of relative risks then the latter would give some expected value in some respect, even though the true expected value is *R / B*. All this hinges, anyhow, on our desire to get or communicate some measure of "overall risk". The present discussion on the contrary focussed on determining cause and the impact of particular events. Hence, an evaluation of overallness is something for somewhere else.

The collected confusions

We identified:

1. "confusus definitionis": mixing up the ETC analysis with other kinds of problems in 2 × 2 × 2 tables.
2. "confusus directionis": not knowing to take either EXZ or EZX.
3. "confusus nomenclaturis": confusing variables and values.
4. "confusus categoriae": taking the wrong category of the right variable as the true cause (new here, for completeness).
5. "confusus magnitudinis": being unsure about the size of the effect (for various reasons).
6. "confusus additionis", a special case of 5: using averages instead of the true parameters.
7. "confusus contributionis": confused about assigning the label "cause" to *f* (which is OK if you are aware of it).
8. "confusus causalitatis", confusing an issue of mere association with an analysis of causality, with different interpretations of the Simpson paradox. (This confusion is a specific kind of 1 and adds flavour to 2 to 7.) (This is not "post hoc ergo propter hoc" since there is no time element yet.)

9. "confusus libertatis": confusing P[*X, Y*] with P[*Y* | *X*] or P[*Y* | *X*] with P[*Y* || *X*] or both.

10. "confusus focis ad risces": focussing on risk and forgetting about safety (and key parameters there).

11. "confusus historiae": using a wrong historical example in trying to clarify a point but thereby actually increasing confusion (see **Appendix B**).

12. "confusus doctoris": getting things wrong because you don't think for yourself but follow your teachers who are confused on some issues and who e.g. focus on neat alphabetical order instead of what they told before. Teachers are liable to tell you what they know and not what you want to known. If you want someone who tries to find out what you want to know in order to explain it, then you need a consultant.

13. "confusus confusi": defining that something can only be a "confounder" if it satisfies relative freedom or conditional independence with the effect measure given the cause (the Cornfield et al. condition). This neglects confounders that show mere statistical association.

14. "confusus conditionis (Simpson et Cornfield et al.)": imposing the Cornfield et al. condition to prevent the Simpson paradox, while a weaker condition is sufficient ($v > 1 - r$). This can also be a special case of the "confusus focis ad risces", since one focusses on a few risk averages while one should use the whole ETC table.

15. "confusus maior": this is not mentioned in the body of the text but supplements the above confusions. To identify a simple cause, we already have the conditions $\{r, b, w, v\} = \{1, 0, 0\ 1\}$. It is not necessary to look for other conditions (such as is done in the "confusus conditionis (S & C)"). Only if we are not speaking about a simple cause but a contributing factor then $\{r, b, w, v\} \neq \{1, 0, 0\ 1\}$; and then we might qualify what kind of factor, e.g. with some condition or not. But such then would would by definition.

It would seem that this taxonomy merely restates what is already very well known to the practical epidemiologists. Yet, some categorization or labelling seems to help understanding the issues. As a cross-over researcher from economics into this $2 \times 2 \times 2$ universe of epidemiology, this author has suffered all these confusions himself at one moment or another. You should not feel ashamed if some happen to you. (But you should feel ashamed right from the start when just follow the analysis in this paper without thinking for yourself.)

Clarity about these angles to confounding would seem to be a prerequisite for handling the more formal approaches of Pearl (1998) or Pearl (2000) chapter 6 to confounding.

PM. On the lighter side there is also the confusion that one doesn't quite know what one is confused about, and on the darker side there is the awkward situation that one thinks that one isn't confused while one is (with the added possibilities that one's environment knows or doesn't know, and everybody else knows or doesn't know).

Conclusions

1-7. See the earlier intermediate conclusions.

8. Causality cannot be resolved with statistical conditions just by themselves. The researcher still needs a model of the world that provides a guide on the direction from cause to effect.

9. But with such a model of the world, the conditions and manipulations discussed here can be used to say more about causality and the effect size.

10. The ETC model is most powerful when we consider a "simple cause" since then we can impose strong conditions on the parameters of the matrix. A scientific discussion on cause and effect gains in clarity if causal chains can be broken down to those. However, models will always refer to "other causes" since it could well be impossible to exclude everything else.

Appendix A: The Simpson paradox

Introduction

The Simpson paradox arises when at least two subpopulations show property A while adding them gives property $\neg A$. Examples and discussions are in Schield (2003), Saari (2001) in voting theory, and Kleinbaum et al. (2003).

Schield (2003) discusses the Simpson paradox and suggests that the Cornfield et al. condition helps to understand it. However, it seems that these are two different issues, at least in the ETC $2 \times 2 \times 2$ world. The Cornfield et al. condition, e.g. $p / q \geq R_F / B_F$ or $R - B \geq R_F - B_F$, concerns the border sum matrices while the Simpson paradox concerns the addition of subpopulations.

Saari (2001) contains a discussion that is targetted specifically at the Simpson paradox without mentioning the Cornfield et al. condition. In Saari's case, the paradox arises merely from the weights of two subpopulations. Translated to our ETC world, the populations F and $\neg F$ have weights f and 1 - f, and the relative risks of both subpopulations would point into one direction while the sum would point into the other direction. This does not necessarily mean that this would be a problem for causality, with its true parameters $\{r, b, w, v\}$, or even that it would be possible given our assumptions.

Creating Simpson paradoxes

The following is a routine to create such a paradox e.g. for treatment-control matrices.

A treatment-control matrix has effect rate p_1 under treatment and effect rate p_2 for the controls.

TreatmentControlMatrix[Set, Pr, n1, p1, n2, p2];

TreatmentControlMatrix[Table]

	Effective	Ineffective	Total
Treatment	n1 p1	n1 − n1 p1	n1
Controls	n2 p2	n2 − n2 p2	n2
Sum	n1 p1 + n2 p2	−p1 n1 + n1 + n2 − n2 p2	n1 + n2

- This is the routine to create a paradox.

? SimpsonParadox

SimpsonParadox[{F1, p1, F2, p2}, {S1, q1, S2, q2}] creates the First and Second (treatment–control) matrices and conditions for the SimpsonParadox, such that the relative risks (or cure rates) are p1/p2 > 1 and q1/q2 > 1 in the subpopulations, while that is precisely the opposite for the total when the data are added. See InequalitySolve when there are numerical values. The joint condition concerns linear combinations of {p1, q1} resp. {p2, q2}, and necessary for the paradox is p2 < p1 < q2 < q1. NB. The sufficient condition is True if the paradox occurs (but only in the > direction

- This is the structure of the problem, with First and Second matrices, and rows 1 and 2.

SimpsonParadox[{n1, p1, n2, p2}, {m1, q1, m2, q2}]

$$\left\{\text{Matrix(1)} \to \begin{pmatrix} \text{n1 p1} & \text{n1} - \text{n1 p1} & \text{n1} \\ \text{n2 p2} & \text{n2} - \text{n2 p2} & \text{n2} \\ \text{n1 p1} + \text{n2 p2} & -\text{p1 n1} + \text{n1} + \text{n2} - \text{n2 p2} & \text{n1} + \text{n2} \end{pmatrix},\right.$$

$$\text{Matrix(2)} \to \begin{pmatrix} \text{m1 q1} & \text{m1} - \text{m1 q1} & \text{m1} \\ \text{m2 q2} & \text{m2} - \text{m2 q2} & \text{m2} \\ \text{m1 q1} + \text{m2 q2} & -\text{q1 m1} + \text{m1} + \text{m2} - \text{m2 q2} & \text{m1} + \text{m2} \end{pmatrix}, \text{Matrix(Sum)} \to \begin{pmatrix} \text{n1 p1} + \\ \text{n2 p2} + \\ \text{n1 p1} + \end{pmatrix}$$

$$\left.\text{Condition} \to \{\text{p1} > \text{p2},\ \text{q1} > \text{q2},\ (\text{m2} + \text{n2})\,(\text{n1 p1} + \text{m1 q1}) < (\text{m1} + \text{n1})\,(\text{n2 p2} + \text{m2 q2})\}\right\}$$

- This is an example where the first two relative risks (or cure rates) are larger than 1 while the sum shows a value lower than 1.

SimpsonParadox[{88, 1/4, 10, 1/5}, {45, 5/9, 90, 4/9}]

$$\left\{\text{Matrix(1)} \to \begin{pmatrix} 22 & 66 & 88 \\ 2 & 8 & 10 \\ 24 & 74 & 98 \end{pmatrix}, \text{Matrix(2)} \to \begin{pmatrix} 25 & 20 & 45 \\ 40 & 50 & 90 \\ 65 & 70 & 135 \end{pmatrix},\right.$$

$$\left.\text{Matrix(Sum)} \to \begin{pmatrix} 47 & 86 & 133 \\ 42 & 58 & 100 \\ 89 & 144 & 233 \end{pmatrix}, \text{Condition} \to \{\text{True, True, True}\}\right\}$$

- This is an example where the three relative risks all point into the same direction.

SimpsonParadox[{88, 1/4, 10, 1/5}, {45, 8/9, 90, 1/3}]

$$\left\{\text{Matrix(1)} \to \begin{pmatrix} 22 & 66 & 88 \\ 2 & 8 & 10 \\ 24 & 74 & 98 \end{pmatrix}, \text{Matrix(2)} \to \begin{pmatrix} 40 & 5 & 45 \\ 30 & 60 & 90 \\ 70 & 65 & 135 \end{pmatrix},\right.$$

$$\left.\text{Matrix(Sum)} \to \begin{pmatrix} 62 & 71 & 133 \\ 32 & 68 & 100 \\ 94 & 139 & 233 \end{pmatrix}, \text{Condition} \to \{\text{True, True, False}\}\right\}$$

Relating the Simpson paradox to the ETC world

Let us consider a proper ETC contingency matrix and extract the submatrices for F and $\neg F$. Since we can always express R and B in terms of w and v, it suffices if we do the discussion in terms of those output parameters, which helps to link up our result to the Cornfield et al. condition.

```
lis = RiskToETCArray[{c, r, b}, {f, q}, {R, B}] // Simplify;
```

```
TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	$(f-1)\,q\,r+c\,R$	$-c\,B+B+b\,(-q\,f+f$
	¬ Confounder	$-(f-1)\,q\,r$	$b\,(f-1)\,(q-1)$
¬ Success	Confounder	$-R\,c+c+q\,(-r\,f+f+r-1)$	$B\,(c-1)-c+f+b\,(f-1)\,(q$
	¬ Confounder	$(f-1)\,q\,(r-1)$	$-(b-1)\,(f-1)\,(q-1)$

Extraction for F.

```
PopF = TransposeToFirst[lis, {3}][[1]];
```

```
TableForm[PopF,
  TableHeadings → {CT[TableHeadings][[1]], CT[TableHeadings][[2]]}]
```

	Cause	¬ Cause
Success	$(f-1)\,q\,r+c\,R$	$-c\,B+B+b\,(-q\,f+f+q-1)$
¬ Success	$-R\,c+c+q\,(-r\,f+f+r-1)$	$B\,(c-1)-c+f+b\,(f-1)\,(q-1)-f\,q+q$

Extraction for $\neg F$.

```
PopNotF = TransposeToFirst[lis, {3}][[2]];
```

```
TableForm[PopNotF,
  TableHeadings → {CT[TableHeadings][[1]], CT[TableHeadings][[2]]}]
```

	Cause	¬ Cause
Success	$-(f-1)\,q\,r$	$b\,(f-1)\,(q-1)$
¬ Success	$(f-1)\,q\,(r-1)$	$-(b-1)\,(f-1)\,(q-1)$

The sum

```
Pop = PopF + PopNotF // Simplify;
```

```
TableForm[Pop,
  TableHeadings → {CT[TableHeadings][[1]], CT[TableHeadings][[2]]}]
```

	Cause	¬ Cause
Success	$c\,R$	$B-B\,c$
¬ Success	$c-c\,R$	$(B-1)\,(c-1)$

Define a general ETCRiskCondition. Since the condition of relative risk > 1 runs the risk of division by zero but is the same as the risk difference > 0, this allows us some freedom.

```
ETCRiskCondition[{{a_, b_}, {c_, d_}}] := ETCRiskCondition[ a / (a + c), b / (b + d)]
```

```
ETCRiskCondition[{{a, b}, {c, d}}]
```

$$\text{ETCRiskCondition}\left(\frac{a}{a+c}, \frac{b}{b+d}\right)$$

Risks for F.

```
RiskPopF = ETCRiskCondition[PopF] // Simplify
```

$$\text{ETCRiskCondition}\left(\frac{(f-1)\,q\,r+c\,R}{c+(f-1)\,q}, \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q}\right)$$

Risks for $\neg F$.

```
RiskPopNotF = ETCRiskCondition[PopNotF ] // Simplify
```

$$\text{ETCRiskCondition}(r, b)$$

The sum

```
RiskPop = ETCRiskCondition[Pop] // Simplify
```

$$\text{ETCRiskCondition}(R, B)$$

Hence for the Simpson paradox:

```
{RiskPopF, RiskPopNotF, RiskPop }
```

$$\left\{\text{ETCRiskCondition}\left(\frac{(f-1)\,q\,r+c\,R}{c+(f-1)\,q}, \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q}\right),\right.$$
$$\left.\text{ETCRiskCondition}(r, b), \text{ETCRiskCondition}(R, B)\right\}$$

A necessary condition for the Simpson paradox in the ETC world is p2 < p1 < q2 < q1, which translates as:

```
RiskPopNotF[[2]] < RiskPopNotF[[1]] < RiskPopF[[2]] < RiskPopF[[1]]
```

$$b < r < \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q} < \frac{(f-1)\,q\,r+c\,R}{c+(f-1)\,q}$$

In case we want a formulation of the latter in terms of the safety parameters:

```
% /. Thread[{R, B} → ETCAverageRisksFromSafety[{c, r, b}, {f, q}, {w, v}]] //
  Simplify
```

$$b < r < 1 - v < 1 - w$$

A sufficient condition for the Simpson paradox in the ETC world is (all would be True for the paradox to occur) (and we use a risk difference because requiring RD > 0 is the same as requiring RR > 1 without the problem of division by zero):

```
{RiskPopF /. ETCRiskCondition[x_, y_] :> (x > y),
  RiskPopNotF /. ETCRiskCondition[x_, y_] :> (x > y),
  RiskPop /. ETCRiskCondition[x_, y_] :> (x < y)}
```

$$\left\{\frac{(f-1)\,q\,r+c\,R}{c+(f-1)\,q} > \frac{B\,(c-1)+b\,(f-1)\,(q-1)}{c+f\,(q-1)-q},\ r>b,\ R<B\right\}$$

In case we want a formulation of the latter in terms of the safety parameters:

```
% /. Thread[{R, B} → ETCAverageRisksFromSafety[{c, r, b}, {f, q}, {w, v}]] //
  Simplify
```

$$\left\{v>w,\ r>b,\ \frac{b\,(-q\,f+f+q-1)-(c+f\,(q-1)-q)\,(v-1)}{c-1}+w+\frac{(f-1)\,q\,(r+w-1)}{c} > 1\right\}$$

Conclusion

The necessary conditions are $b < r < 1 - v < 1 - w$, which translates too as $w < v < 1 - r$. Crucially, for causality we will tend to impose $v > 1 - r$ (with the extreme values $1 > 0$). Ssufficient conditions for occurrence (not prevention) that are identified are (by definition): (1) $v > w$, (2) $r > b$, (3) $R < B$. The combination is not likely to occur given the assumptions that we have formulated for a serious causal model. $R < B$ would not occur if $v > 1 - r$. If we would formulate the paradox in a reverse direction, then we would have to assume $r < b$, which neither makes sense for a causal model. See further the body of the text for the summary conclusion.

Note

We can employ above expression to also give the "Adjusted Relative Risk", also produced by the routine ETCAdjustedRRisk.

```
{R / B, RiskPopNotF, RiskPopF, f * RiskPopF + (1 - f) * RiskPopNotF} /.
    ETCRiskCondition[x_, y_] :> (x / y) // Simplify
```

$$\left\{\frac{R}{B},\ \frac{r}{b},\ \frac{(c+f\,(q-1)-q)\,((f-1)\,q\,r+c\,R)}{(B\,(c-1)+b\,(f-1)\,(q-1))\,(c+(f-1)\,q)},\right.$$
$$\left.\frac{r-f\,r}{b}+\frac{f\,(c+f\,(q-1)-q)\,((f-1)\,q\,r+c\,R)}{(B\,(c-1)+b\,(f-1)\,(q-1))\,(c+(f-1)\,q)}\right\}$$

% // Variables

$$\{b, B, c, f, q, r, R\}$$

%% /. Thread[{R, B} → ETCAverageRisksFromSafety[{c, r, b}, {f, q}, {w, v}]] // Simplify

$$\left\{\frac{(c-1)\,(c\,(w-1)+(f-1)\,q\,(r+w-1))}{c\,(b\,(f-1)\,(q-1)+(c+f\,(q-1)-q)\,(v-1))},\ \frac{r}{b},\ \frac{w-1}{v-1},\ \frac{r\,(-v\,f+f+v-1)+b\,f\,(w-1)}{b\,(v-1)}\right\}$$

% // Variables

$$\{b, c, f, q, r, v, w\}$$

Appendix B: Fisher on smoking and confounding

Statement

The author is an independent researcher and has no material, political or moral interests in any issue on smoking and lung cancer.

Introduction

The following example and discussion is based upon Schield (2003), "Simpson's paradox and Cornfield's conditions". Shield presents a historical case of a discussion on the possible causes of lung cancer, with two "Letters to the editor" by Sir R.A. Fisher (1958ab). Historical examples often help clarifying an issue indeed. Yet, in this case it appeared, at least to this author, after some struggle for clarity, that the use of this historical example actually contributed to confusion. That is, it would contribute if we were to use it in the main body of the text. There was a version of this paper that actually proceeded in this manner. But it totally confused both the history and the subject. Thus, here in the appendix, the historical issue can find a good place for proper sub-discussion, and then it will be clarifying again. Schield apparently followed the history as it afterwards got to be told yet the objective of his paper is different from our objectives. For us, it appears that the history needs to be rewritten. So, this is a new subject. History on Fisher and smoking needs to be rewritten.

An example problem setting

In 1958 it was for the first time seriously conjectured that smoking caused lung cancer. Fisher apparently had his doubts, though. Fisher (1958ab) considered fraternal and identical twins and compared their habits of smoking, that he categorized into being either alike or unlike. He found that 51% of male fraternal twins and 24% of male identical twins had distinct different habits in smoking (smoking versus non-smoking, or cigarette versus pipe). This indicated a strong effect of genetics and thus created the possibility, also in this debate on the cause for lung cancer, that smoking was just a confounder. Genotype might very well cause both lung cancer and a disposition to smoke. Fisher's data on smoking habits were the following (correcting a typesetting error):

CT[Set, "Fisher Twins"]

		Fraternal	Identical
Alike	Male	15	39
	Female	9	44
Unlike	Male	16	12
	Female	9	9

Fisher (1958a): "(...) of the (male) dizygotic pairs (...) 16 out of the 31 are distinctly different, this being 51 per cent. as against 24 per cent. (...) among the monozygotic. the (male) monozygotic twins show closer similarity and fewer divergences than the dizygotic. There can therefore be little doubt that the genotype exercises a considerable influence on the smoking and on the particular habit of smoking adopted (...)".

Note that the discussion quickly becomes complicated. These data don't tell anything about getting cancer or the amount of smoking. They just show equality of habit. The "alike" groups would be split over smoking or not. So there is only a suggestion of the influence of the genotype. Fisher definitely does not say anything particular about the prevalences of hidden genetic factors that would be a common cause for both smoking and cancer, and neither does he say anything about the rates of risk. Fisher merely pointed to genetics as a common factor that should not be overlooked when establishing causality for an important disease. In the end, it indeed will also be molecules that interact with molecules.

Nevertheless, the ratio's of 51% and 24% struck a chord, and got interpreted as such prevalences.

Cornfield et al. 1959 stated: “if, cigarette smokers have 9 times the risk of nonsmokers for developing lung cancer, and this is not because cigarette smoke is a causal agent, but only because cigarette smokers produce hormone X, then the proportion of hormone-X-producers among cigarette smokers must be at least 9 times greater than that of non-smokers. If the relative prevalence of hormone-X-producers is considerably less than ninefold, then hormone X cannot account for the magnitude of the apparent effect.” (Taken from Schield (2003).)

They thus identified a ‘minimum effect size’ for possible confounders: the relative prevalence (p versus q) must be at least the seeming relative risk (R_F versus B_F). One can impose this condition if there is adequate (theoretical) reason that the causal relations should be proportional. This is OK as it is.

Fisher had mentioned the numbers of 51% and 24% and this got interpreted as a 2-fold relative prevalence. This could not explain the 9-fold relative risk for smoking itself.

But this was not what Fisher had expressed. Thus, the minimum effect size got into the literature by confusion. Schield (2003) states: “Fisher never replied.” But would you “reply” if you say “A” and somebody else says “not B but C” ? Fisher already had replied on confusion, and one can imagine that he would have been perplexed when this very remark targetted towards clarity was confounded itself.

The Fisher model

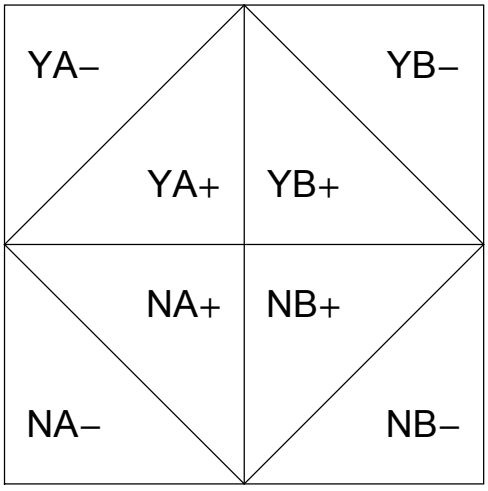
Fisher’s suggestion comes down to creating groups with different genotypes. A “genotype” can be defined such that the risk for getting cancer and the inclination to smoke are biologically proportional. This is modelled as statistical independence. It is crucial to see that biological proportionality is modelled such - and the biological explanation is that the genotype here deals with “other causes”. If the risk of getting cancer and the inclination to smoke are not statistically independent then we split up the group again, if necessary down to the individual level. For ease, we take two groups, those with “riskier” genes and those with “safer” genes. Risk group A of size n_A has risk S_A to get cancer (a “Success”) and inclination F_A to smoke. Background risk group B has n_B, S_B and F_B. By necessity of our definition, in group A the probabilities $\{S_A, 1 - S_A\}$ and $\{F_A, 1 - F_A\}$ would be statistically independent. The non-smokers in A would still get cancer at rate S_A. Each group has a proportional relation to its two variables that does not change with group size. The overall outcome is the sum of those tables weighed by the sizes of the groups.

```
TableForm[Transpose[{ n_A PrTable[S_A, F_A], n_B PrTable[S_B, F_B]}, {2, 1, 3}],
  TableHeadings →
   {{"Cancer", Not["Cancer"]}, {"A", "B"}, {"Smoking", Not["Smoking"]}}]
```

		A	B
Cancer	Smoking	$F_A\, n_A\, S_A$	$F_B\, n_B\, S_B$
	¬ Smoking	$(1-F_A)\, n_A\, S_A$	$(1-F_B)\, n_B\, S_B$
¬ Cancer	Smoking	$F_A\, n_A\, (1-S_A)$	$F_B\, n_B\, (1-S_B)$
	¬ Smoking	$(1-F_A)\, n_A\, (1-S_A)$	$(1-F_B)\, n_B\, (1-S_B)$

In this square Y / N stands for having (risk for) cancer and the + / - stand for the presence of the confounder. The cause would be the genotype which thus occupies the columns.

```
ETCSquare[Label, First → {"Y", "A", "+"}, Last → {"N", "B", "−"}];
```



For this dichotomous and stratefied model (with possibly more strata than just two), Kleinbaum et al. (2003:423) advise: “The Mantel-Haenszel test is the most widely used and recommended procedure for testing an overall association in a stratefied analysis.” But they add (p429), pointing to a Simpson effect: “When there is opposite direction interaction, use of the Mantel-Haenszel test is often inappropriate because it may mask a strong interaction effect that reflects the true exposure disease relationship.” The latter point is only noted here. Our reason to quote this is merely to point out that this Fisher model is do-able and does in no way relate to criticism of the minimum effect size issue that history attributes to it.

The Fisher model is special in that it assumes (1) statistical independence, (2) proportionality to group sizes, and (3) a degree of presence (different rates per group, all down the alphabet, since Fisher did not limit himself to two groups), (4) the genotype is

a common cause for both confounder and effect. Alternatives to these assumptions are: (ad 1) statistical dependence, for example due to random causes not mentioned, or that we have not disaggregated enough, (ad 2) non-proportionality for various other reasons than aggregation, (ad 3) presence or absence (thus dichotomy only), (ad 4) allow causal independence between cause and confounder.

The analysis in the main body of the text departs from the Fisher model in these ways: (ad 4) allow causal independence between cause and confounder, (ad 3) presence or absence, (ad 1) allow statistical dependence, (ad 2) allow indeterminacy on proportionality, since we have only one table. We don't have "groups" but subpopulations who are exposed to the cause or not. Of necessity these subpopulations are groups anyhow, yet they are not the groups that are by definition intended to have independent probabilities.

If the cause is absent then one generally would presume that the effect does not occur, yet we still will allow for some "background risk" due to other causes (" causes not mentioned"). Or alternatively, if there should always be a value 0 somewhere, then one would have to reshuffle the data.

By consequence, the main body of the text uses different variable names than the one in this appendix on Fisher's model.

It must also be remarked that the Fisher model does not say anything about causality. The direction of events derives from a different kind of reasoning than the mere way of tabulating the data.

Some other remarks

The $2 \times 2 \times 2$ case that we studied in the main body of the text is sufficiently general to be used to study the Fisher model as well, since that is minimally $2 \times 2 \times 2$ as well. Yet one must keep those issues in mind when making the translation ("confusus definitionis"). An inquisitive reader may note that our example table, that is entirely fictitious, already contains $p = 51\%$ and $q = 24\%$, so that the 800 persons could be seen as fraternal twins, and the confounder would be "having distinctly different smoking habits". Also there is a relative average risk of about 9. The other numbers on effect and cause are entirely fictitious though. It thus is dubious what one can do with this, in particular since "having a different smoking habit than your twin brother" is non-informative on smoking at all.

Needless to say, this author thinks that smoking is a cause for cancer. Yet, it would not be right to misread Fisher. His warning was against confounding, about mistaking correlation for causation. His model was adequate and his observation of the influence of genetics was important. In the same way of "getting the record straight", it must be remarked that, though Schield (2003) follows the historically grown interpretation that Fisher's 51% and 24% would be prevalences, which they are not, this author is enormously indebted to Schield for helping to understand the issue. It did take some effort to first understand Schield (2003) and then to see that the issue is slightly different, yet, in the end it was an important point to start from.

It still is not entirely clear to this author whether this issue is one of proportionality or one of imposing conditional independence. For epidemiologists it may be natural to think in terms of conditional independence, so that for them the insight of Cornfield et al. derives from the proportionality. For an economist though, contingency tables in principle might take any values, everything is proportional to n, and the insight is from imposing relative freedom.

It is not clear to this author whether Fisher historically knew about the notion of conditional independence. It would be strange however to assume that he would not, in all practical matters, have used the notion. Perhaps the notation and the developed mathematics must have been unknown since these were of a later date, yet, those concepts, to an important degree, merely express common sense. It would seem, at least to this author, that Fisher might have thought: 'You folks may come with all kinds of explanations for lung cancer, yet, please be aware that correlation is no causation, and for example that new wonderful invention of genetics might be the true cause'. Which thought is caught by that double bar in P[Y || X].

Epidemiological language and conventions

This sub-section uses Schields notation. For translation (us = Schield): S = E, C = C, F = A.

Schield (2003:3) in the first column states: "If factor A (smoking) had no effect on the likelihood of an observable effect E (lung cancer), Cornfield et al. proved that the prevalence of the actual cause (C) must satisfy: P(C|A) / P(C|A') > P(E|A) / P (E|A')." (In our notation this is $p / q > R_F / B_F$.)

However, this statement is totally incomprehensible since it has not been defined for the reader what "had no effect" means. It might be (1) P[E, A] or (2) P[E | A] or (3) P[E | C,

A] or (4) P[E || A]. Schield (2003:3) then in the second column gives a sufficient condition for "no effect", which is conditional independence, of E ⊥ A | C. It is not clear whether sufficiency is actually the definition. Also, it seems to amount to begging the question. Yes, if there is this relative freedom then the Cornfield et al. condition holds, as well as $R - B \geq R_F - B_F$, but it does not of necessity follow that A (our *F*) is a confounder *if and only if* E ⊥ A | C.

Schield (2003:3) states: "The necessary condition of Cornfield et al is the positive side of Simpson's Paradox. It allowed statisticians to conclude that, to the best of their knowledge, smoking caused cancer – based on observational studies."

This is not quite true. The main body of the text showed that the EXZ analysis can be inverted to some EZX form that satisfies all conditions. One requires additional theory on how the world operates to truly arrive at a decision.

In his own Appendix, Schield (2003:7) quotes the Appendix A of Cornfield et al. (who use a different notation again): "Let the disease rate for those exposed to the causal agent *B*, be r_1 and for those not exposed, r_2, each rate being unaffected by exposure or nonexposure to the noncausal agent, A."

However, this statement is totally incomprehensible since it has not been defined for the reader what "being unaffected by exposure or nonexposure" means. It is only fortunate that the present author heard about the notion of "conditional independence" and was able to make an educated guess. But your author remained and remains perplexed why one would make such an assumption, since it might very well be that $r \neq R$. These are serious diseases. One should not impose assumptions from thin air and without even speaking about them. But the author has had no boot camp in epidemiology, so he may have missed basic training. Part of the solution appears the paradigm that biological proportionality tends to be modelled as statistical independence. And it is likely the simplest model that one tries. Perhaps that is all.

Appendix C: Deductions on safety

It will be useful to substitute:

```
sol = Solve[c == p f + q (1 – f), p][[1]]
```

$$\left\{p \to \frac{c + f\,q - q}{f}\right\}$$

This is the total when S and $\neg S$ are added.

```
ETCTable["TC", f, {p, q}]
```

	Cause	¬ Cause	Total
Confounder	$f\,p$	$f\,(1-p)$	f
¬ Confounder	$(1-f)\,q$	$(1-f)\,(1-q)$	$1-f$
Sum	$f\,p+(1-f)\,q$	$-f\,p+f\,q-q+1$	1

```
mat = Take[%, 2, 2];
```

This gives the part for $\neg S$.

$$w = \text{P}[\neg S \mid C, F] = \text{P}[\neg S, C, F] / \text{P}[C, F] = \text{P}[\neg S, C, F] / (p\ f)$$

$$v = \text{P}[\neg S \mid \neg C, F] = \text{P}[\neg S, \neg C, F] / \text{P}[\neg C, F] = \text{P}[\neg S, \neg C, F] / ((1 - p)\ f)$$

$$e = \text{P}[\neg S \mid C, \neg F] = \text{P}[\neg S, C, \neg F] / \text{P}[C, \neg F] = \text{P}[\neg S, C, \neg F] / (q\ (1 - f))$$

$$a = \text{P}[\neg S \mid \neg C, \neg F] = \text{P}[\neg S, \neg C, \neg F] / \text{P}[\neg C, \neg F] = \text{P}[\neg S, \neg C, \neg F] / ((1 - q)\ (1 - f))$$

```
{{P[¬ S, C, F], P[¬ S, ¬ C, F]}, {P[¬ S, C, ¬ F], P[¬ S, ¬ C, ¬ F]}} ==
    {{w p f, v (1 – p) f}, {(1 – r) q (1 – f), (1 – b) (1 – q) (1 – f)}} /. sol
```

$$\begin{pmatrix} P(\neg S, C, F) & P(\neg S, \neg C, F) \\ P(\neg S, C, \neg F) & P(\neg S, \neg C, \neg F) \end{pmatrix} = \begin{pmatrix} (c + f\,q - q)\,w & f\left(1 - \frac{c+f\,q-q}{f}\right)v \\ (1-f)\,q\,(1-r) & (1-b)\,(1-f)\,(1-q) \end{pmatrix}$$

This has in fact be put in a separate routine as well.

```
ETCTable["TC|–S", {c, r, b}, {f, q}, {w, v}]
```

	Cause	¬ Cause
Confounder	$(c-(1-f)\,q)\,w$	$f\left(1-\frac{c-(1-f)\,q}{f}\right)v$
¬ Confounder	$(1-f)\,q\,(1-r)$	$(1-b)\,(1-f)\,(1-q)$
Sum	$(1-f)\,q\,(1-r)+(c-(1-f)\,q)\,w$	$(1-b)\,(1-f)\,(1-q)+f\left(1-\frac{c-(1-f)\,q}{f}\right.$

```
matnots = Take[%, 2, 2]
```

$$\begin{pmatrix} (c-(1-f)\,q)\,w & f\left(1-\frac{c-(1-f)\,q}{f}\right)v \\ (1-f)\,q\,(1-r) & (1-b)\,(1-f)\,(1-q) \end{pmatrix}$$

Substraction from the total gives the part for *S*.

```
mats = mat – matnots /. sol // Simplify
```

$$\begin{pmatrix} -(c+(f-1)\,q)\,(w-1) & (c+f\,(q-1)-q)\,(v-1) \\ -(f-1)\,q\,r & b\,(f-1)\,(q-1) \end{pmatrix}$$

And this is the whole matrix again

```
matsol = {mats // Transpose, matnots // Transpose} // Simplify;

XminusAToAminusX[%];

TableForm[%]
```

$$\begin{array}{ll} (c-(1-f)\,q)\,(1-w) & -(c-f\,(1-q)-q)\,(1-v) \\ (1-f)\,q\,r & b\,(1-f)\,(1-q) \\ (c-(1-f)\,q)\,w & (-c+f\,(1-q)+q)\,v \\ (1-f)\,q\,(1-r) & (1-b)\,(1-f)\,(1-q) \end{array}$$

A small check:

```
Add[%] // Simplify
```

1

Hence, the routine creates that output.

```
lis = SafetyToETCArray[{c, r, b}, {f, q}, {w, v}];

TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	$(c-(1-f)\,q)\,(1-w)$	$(-c+f\,(1-q)+q)\,(1-v)$
	¬ Confounder	$(1-f)\,q\,r$	$b\,(1-f)\,(1-q)$
¬ Success	Confounder	$(c-(1-f)\,q)\,w$	$(-c+f\,(1-q)+q)\,v$
	¬ Confounder	$(1-f)\,q\,(1-r)$	$(1-b)\,(1-f)\,(1-q)$

Appendix D: Deductions on risk

Recall:

ETCTable["TC", f, {p, q}]

	Cause	¬ Cause	Total
Confounder	$f\ p$	$f\ (1-p)$	f
¬ Confounder	$(1-f)\ q$	$(1-f)\ (1-q)$	$1-f$
Sum	$f\ p+(1-f)\ q$	$-f\ p+f\ q-q+1$	1

ETCTable["ET", c, {R, B}]

	Cause	¬ Cause	Total
Success	$c\ R$	$B\ (1-c)$	$B\ (1-c)+c\ R$
¬ Success	$c\ (1-R)$	$(1-B)\ (1-c)$	$c\ B-B-c\ R+1$
Sum	c	$1-c$	1

Instead of these averages we are interested in the driving risks (using above TC table):

r = P[S | C, ¬F] = P[S, C, ¬F] / (P[C | ¬F] P[¬F]) = P[S, C, ¬F] / (q (1 - f))

b = P[S | ¬C, ¬F] = P[S, ¬C, ¬F] / (P[¬C | ¬F] P[¬F]) = P[S, ¬C, ¬F] / ((1 - q) (1 - f))

And this allows us to understand what happens in general when the confounder is not present:

P[S, C | ¬F] = P[S, C, ¬F] / P[¬F] = $r\ q$

P[S, ¬C | ¬F] = P[S, ¬C, ¬F] / P[¬F] = b (1 - q)

When we consider the group ¬F as a whole, conditionally, then we find the following table - which is also the second ratio table printed in the above ETCStatistics output.

- This looks only at the group with ¬F. All values must be multiplied by 1 - f.

ETCTable["ET", q, {r, b}]

	Cause	¬ Cause	Total
Success	$q\ r$	$b\ (1-q)$	$b\ (1-q)+q\ r$
¬ Success	$q\ (1-r)$	$(1-b)\ (1-q)$	$q\ b-b-q\ r+1$
Sum	q	$1-q$	1

If we substract this result (multiplied by 1 - f) from the earlier total, we get the matrix for the group with F. Hence we have parameterized the whole 2 × 2 × 2 matrix.

lis = RiskToETCArray[{c, r, b}, {f, q}, {R, B}];

```
TableForm[lis, TableHeadings → CT["ETC", TableHeadings]] // Simplify
```

		Cause	¬ Cause
Success	Confounder	$(f-1)\,q\,r+c\,R$	$-c\,B+B+b\,(-q\,f+f$
	¬ Confounder	$-(f-1)\,q\,r$	$b\,(f-1)\,(q-1)$
¬ Success	Confounder	$-R\,c+c+q\,(-r\,f+f+r-1)$	$B\,(c-1)-c+f+b\,(f-1)\,(q$
	¬ Confounder	$(f-1)\,q\,(r-1)$	$-(b-1)\,(f-1)\,(q-1)$

```
lis2 = SafetyToETCArray[{c, r, b}, {f, q}, {w, v}];
```

```
TableForm[lis2, TableHeadings → CT["ETC", TableHeadings]] // Simplify
```

		Cause	¬ Cause
Success	Confounder	$-(c+(f-1)\,q)\,(w-1)$	$(c+f\,(q-1)-q)\,(v-1)$
	¬ Confounder	$-(f-1)\,q\,r$	$b\,(f-1)\,(q-1)$
¬ Success	Confounder	$(c+(f-1)\,q)\,w$	$(-c+f-f\,q+q)\,v$
	¬ Confounder	$(f-1)\,q\,(r-1)$	$-(b-1)\,(f-1)\,(q-1)$

```
eqs = Thread[Flatten[lis] == Flatten[lis2]] /. True :> Sequence[];
```

```
sol = Solve[eqs, {R, B}] // Simplify // XminusAToAminusX
```

$$\left\{\left\{R \to -\frac{(1-f)\,q\,(-r-w+1)}{c} - w + 1,\ B \to \frac{b\,(f-1)\,(q-1)+(-c+f\,(1-q)+q)\,(1-v)}{1-c}\right\}\right\}$$

Hence the routine reproduces that.

```
rs = ETCAverageRisksFromSafety[{c, r, b}, {f, q}, {w, v}]
```

$$\left\{-w+\frac{(1-f)\,q\,(r+w-1)}{c}+1,\ \frac{b\,(f-1)\,(q-1)+(-c+f\,(1-q)+q)\,(1-v)}{1-c}\right\}$$

In case c = q: R == r (1 - f) + f (1 -w), B == b (1 - f) + f (1 - v):

```
{R == r (1 − f) + f (1 − w), B == b (1 − f) + f (1 − v)} /. sol[[1]] /. c → q // Simplify
```

{True, True}

PM. The risk parameterization using safety simplifies to the one using risk, when we substitute the risks created from safety.

```
lis = SafetyToETCArray[{c, r, b}, {f, q}, {w, v}];

lis2 = RiskToETCArray[{c, r, b}, {f, q}, rs];

lis == lis2 // Simplify
```

True

Appendix E: The risk difference

Schield (2003) provides evidence that the risk difference would be instructive to identify a Simpson paradox, not only algebraically but also psychologically.

Schield (2003:5): "Consider two hospitals: a city hospital and a rural hospital. The death rate is 3% of cases at the city hospital versus 2% at the rural. The combined death rate is 2.7%. Thus, it seems that the rural hospital is safer than the city hospital. (...) Now consider a plausible confounding factor: the condition of the patient's health. We find that overall the death rate among patients in poor condition is 3.8% while that among patients in good condition is 1.2%. Here the simple difference in death rates by patient condition (2.6 percentage points) is greater than the simple difference in death rates by hospital (1 percentage point). Thus we have strong reason to be concerned about a possible Simpson's Paradox reversal of the association between hospital and death rate. To guard against such a reversal we can take into account (control for) patient condition when comparing the death rates for these two hospitals."

As we have seen, the mathematics of the ETC problem is fairly simple, but translation remains a stumble block. In above story, the true cause is the health condition while Schield labels it the "confounder". So we would first translate the text to the EZX situation and then invert to EXZ again. Let us try to do this in one step. The second element in the translation is that the discussion is indiscriminate about "death rates" while some are marginals (averages) while other might be parameters. There are no clearly stated marginals *s, c* and *f* so we have to infer those from the averages. The third snitch is that epidemiologists assume $\{R, B\} = \{r, b\}$, for otherwise they could not keep the number of variables down. The fourth problem is that Schield refers to the Simpson paradox and uses Cornfield's condition $R - B \geq R_F - B_F$ to solve it, which is needlessly strong. A fifth point is that the Simpson paradox would not occur in a truly causal model. Admittedly, though, the distinction between city and rural hospitals is hardly a causal one, so that this is indeed a model of aggregating subpopulations, which is the habitat of the Simpson paradox. A sixth issue is that "to control" is not really defined so that it is not really clear what the solution is. If we had the solution then it would also become clear what the original problem was (what we lacked in knowledge).

Translating we find S = death, C = bad health, F = city hospital, $s = 2.7\%$, $R_F = 3\%$, $B_F = 2\%$, $r = R = 3.8\%$, $b = B = 1.2\%$. From the section above on the relative freedom of

the variables we know that $R_F = p\ r + (1 - p)\ b$ and $B_F = q\ r + (1 - q)\ b$. There is only one way that these numerical values can fit our equations.

```
vals = {s → 0.027, r → 0.038, b → 0.012, Rf → .03, Bf → .02};
eqs = { s == c R + (1 – c) B,
        s == f Rf + (1 – f) Bf,
        c == p f + q (1 – f),
        Rf == p r + (1 – p) b } /. {R → r, B → b} /. vals
```

$$\{0.027 = 0.012\,(1 - c) + 0.038\,c,\ 0.027 = 0.02\,(1 - f) + 0.03\,f,\ c = f\,p + (1 - f)\,q,\ 0.03 = 0.012\,(1 - p) + 0.038\,p\}$$

```
sol = Solve[eqs, {c, f, p, q}]
```

$$\{\{c \to 0.576923,\ p \to 0.692308,\ q \to 0.307692,\ f \to 0.7\}\}$$

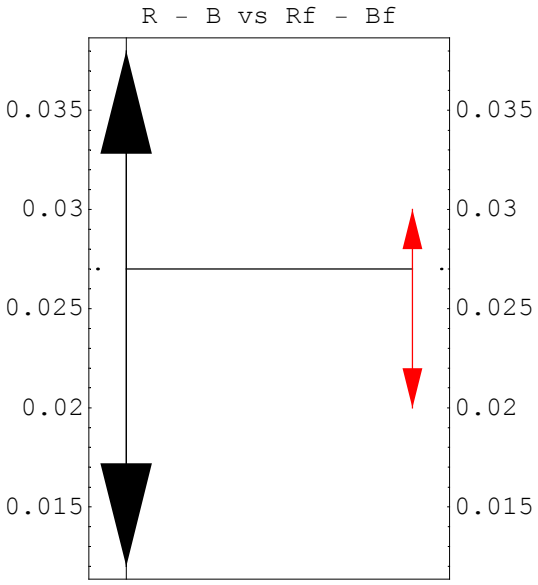
```
lis = RiskToETCArray[{c, r, b}, {f, q}, {r, b}] /. vals /. sol[[1]];

TableForm[lis, TableHeadings → CT["ETC", TableHeadings]] /.
  {"Cause" → "Bad health",
    "Confounder" → "City hospital", "Success" → "Death"}
```

		Bad health	¬ Bad health
Death	City hospital	0.0184154	0.00258462
	¬ City hospital	0.00350769	0.00249231
¬ Death	City hospital	0.4662	0.2128
	¬ City hospital	0.0888	0.2052

The reason why we took Schield's example is to follow his suggestion on the risk difference. The Schield plot shows the differences between the four parameters with the total average at the center of gravity. Since the risk difference of the cause (bad health) is larger than the risk difference of the seeming cause (city hospital), the Cornfield condition is satisfied. Indeed, if you had a reverse state of mind, and thought that the true cause was the difference in quality in city and rural hospitals, then you would have to worry about confounding and the Simpson paradox. Our analysis shows that we should not stick to just these risk rates but consider the whole ETC 2 × 2 × 2 table. The assumption that $\{R, B\} = \{r, b\}$ is too quick and likely hides the true relations. (Thus there is no reason to further investigate the issue of the relative performance of the hospitals.)

- $\{R, B\} = \{r, b\}$

```
SchieldShow[0.027, 0.038, 0.012, .03, .02, "R – B vs Rf – Bf"];
```



Appendix F: A counterfactual in Pearl (2000)

Pearl (2000:35-36) gives a wonderful example on counterfactuals. The problem situation differs from our ETC world since this example may have two causes. But a short discussion will clarify both the ETC assumptions and a bit about the counterfactual. Let us first re-create the problem and then summarize our finding. There are two random variables u_1 and u_2 that can take values {1, 0} with a flip of a coin (P = 1/2). The variables are S = death, $C = u_2$, $F = u_1$.There are two models:

Model 1: $S = C$

Model 2: $S = C\ F + (1 - C)\ (1 - F)$

The reader will note that Pearl (2000) calls $F = u_1$ the "treatment" variable, which suggests that it would be the cause. But from the structure of Model 1 we can see that u_2 is the cause. Perhaps one might better call u_2 the "treatment". Thus it seems that this

example on counterfactuals is a bit confounded with treating a EZX model instead of a EXZ model. To allow the comparison with Pearl (2000) we stick to his label $F = u_1$ = "treatment".

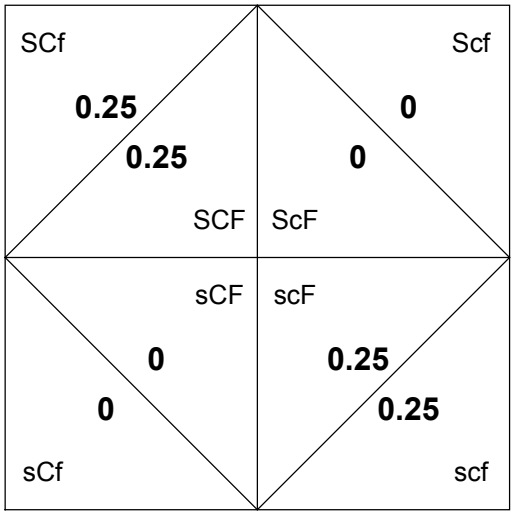
```
lab = {"Success" → "Death", "Cause" → "u2" , "Confounder" → "Treatment"};
```

- This is Model 1. It is a perfect simple causal model.

```
func[x_, y_, z_] := If[x == y, .25, 0]

lis = Outer[func, {1, 0}, {1, 0}, {1, 0}];

TableForm[lis, TableHeadings → CT["ETC", TableHeadings]] /. lab
```

		u_2	$\neg u_2$
Death	Treatment	0.25	0
	¬ Treatment	0.25	0
¬ Death	Treatment	0	0.25
	¬ Treatment	0	0.25

```
ETCSquare[lis];
```



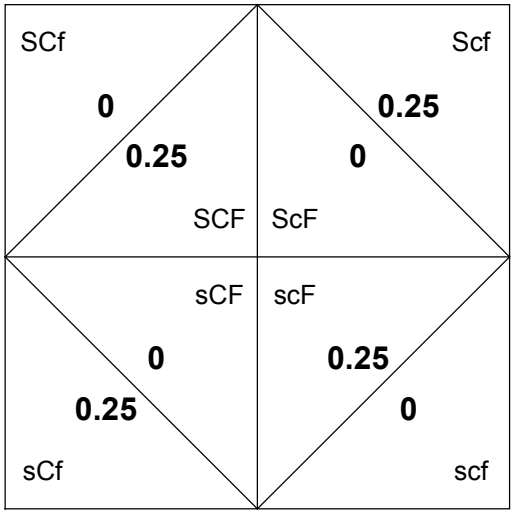
- This is Model 2.

```
func2[x_, y_, z_] := If[x == yz + (1 − y) (1 − z), .25, 0]

lis2 = Outer[func2, {1, 0}, {1, 0}, {1, 0}];
```

TableForm[lis2, TableHeadings → CT["ETC", TableHeadings]] /. lab

		u_2	$\neg u_2$
Death	Treatment	0.25	0
	¬ Treatment	0	0.25
¬ Death	Treatment	0	0.25
	¬ Treatment	0.25	0

ETCSquare[lis2];



One may check that both matrices have the same marginal distribution {{.25, .25}, {.25, .25}} when $C = u_2$ or $F = u_1$ is summed out. Thus the relative risks for both cause and confounder are 1.

Pearl (2000:36): "Model 1 corresponds to treatment (X) that has no effect on any of the subjects; in model 2, every subject is affected by treatment. The reason that the two models generate the same distribution is that model 2 describes a mixture of two subpopulations. In one ($u_2 = 1$), each subject dies ($y = 1$) if and only if treated; in the other ($u_2 = 0$), each subject recovers ($y = 0$) if and only if treated."

Pearl is interested at this point in the counterfactual. For lack of a better notation we can write $Q = P[(y = 0 \mid x = 0) \mid (y = 1, x = 1)]$ for the counterfactual that a person who died under treatment (the outer condition) would have recovered under non-treatment (the inner condition).

Pearl (2000:36): "The value of Q differs from these two models. In model 1, Q evaluates to zero, because subjects who died correspond to $u_2 = 1$ and, since the treatment has no effect on y, changing X from 1 to 0 would still yield $y = 1$. In model 2, however, Q

evaluates to unity, because subjects who died under treatment must correspond to $u_2 = 1$ (i.e. those who die if treated), meaning they would recover if and only if not treated."

We can capture the counterfactual also in the basic statistics. This corresponds with Pearl's remark: "knowledge about the actual process behind P($y \mid x$) is needed for the computation". Once we have the full crosstable with all 8 entries then we have full information about the table, and then also the counterfactuals can be calculated. But it is not entirely clear why we should link this issue to counterfactuals. Having the full table allows us to calculate all kinds of things, including counterfactuals. But we want the full table for various purposes, not only counterfactuals. To put disproportionate emphasis on counterfactuals might be a "confusus counterfactualis".

When we run the statistics then we get (using Hold[SafetyToETCArray] to identify the parameters):

```
SafetyToETCArray[] /. ETCStatistics[lis, Print → False]
```

```
Hold[SafetyToETCArray][{0.5, 1., 0}, {0.5, 0.5}, {0, 1.}, 1.]
```

```
SafetyToETCArray[] /. ETCStatistics[lis2, Print → False]
```

```
Hold[SafetyToETCArray][{0.5, 0, 1.}, {0.5, 0.5}, {0, 1.}, 1.]
```

Thus:

- Model 1 is a perfect model for a simple cause, with $r = 1$, $b = 0$, $w = 0$ and $v = 1$.
- Model 2 describes a *perverse* simple cause, albeit with $w = 0$ and $v = 1$, but $r = 0$, $b = 1$. The background risk determines all and the "cause" is totally ineffective. Though u_2 was correctly identified as the cause for Model 1 we mistakenly believed, purely on its name, that we have a related kind of causal structure in Model 2. Indeed, we have the same u_2 and we also see deaths, so, there must be something constant. But the only thing that is constant is "there is a probability distribution". What we did not notice is that between Model 1 and Model 2 there was a huge shift in the actual probabilities, notably on the risks for the cause and the background. It is like buying a coke because of the label on the bottle but then discovering that someone had changed the (formula for the) drink.

- We may try to do the EZX transformation to see whether Treatment now is the true cause. But transponation gives us the same data format.

```
CT[Set, Label → "Pearl2",
    Dimensions → CT[Dimensions], Data → lis2, Source → "Created"];
```

```
CT[Order, {"Effect", "Confounding"}, "Pearl2"] /. lab
```

		Treatment	¬ Treatment
Death	u_2	0.25	0
	¬ u_2	0	0.25
¬ Death	u_2	0	0.25
	¬ u_2	0.25	0

- Indeed, also if we try Treatment as the cause then we still have the perverse $r = 0$ and $b = 1$.

```
SafetyToETCArray[] /. ETCStatistics[%, Print → False]
```

```
Hold[SafetyToETCArray][{0.5, 0, 1.}, {0.5, 0.5}, {0, 1.}, 1.]
```

- A solution is to call absence of the cause to be the true cause, with $r = 1$ and $b = 0$ though at a price of $w = 1$ and $v = 0$.

```
CT[Switch, "Pearl–Invert", "Pearl2", "Truth" → {Not["Cause"], "Cause"}]
```

		¬ Cause	Cause
Success	Confounder	0	0.25
	¬ Confounder	0.25	0
¬ Success	Confounder	0.25	0
	¬ Confounder	0	0.25

```
SafetyToETCArray[] /. ETCStatistics[%, Print → False]
```

```
Hold[SafetyToETCArray][{0.5, 1., 0}, {0.5, 0.5}, {1., 0}, 1.]
```

A strong conclusion would be that Pearl's Model 2 may very well be a possible model with nice probabilities in a 2 × 2 × 2 table, but it does not fit the ETC mold. For ETC we have been considering a cause and a confounder, but in Model 2 we meet with two causes, and we have not modelled such a case including their interaction. We may have suffered a bit from the "confusus definitionis", mistaking a 2 × 2 × 2 table for the ETC problem, just because it was a 2 × 2 × 2 table.

It may well be that Pearl's comment on the subpopulations is relevant. That would introduce a fourth variable. It all depends upon the problem. The fourth variable should explain why the same cause from Model 1 should suddenly get perverse effects.

But another solution is to stick to the three-variable world and pay closer attention to the actual definition of the probability distribution for Model 2. There we find that there is only a success if we have, in logical terms, the true cause $\mathbb{C} = (C \wedge F) \vee (\neg C \wedge \neg F)$. Thus instead of going into the fourth dimension we actually have a smaller model from the 2×2 world with a perfect simple causal structure. And there are no subpopulations to speak of, since those have not been specified, and without such specification all probabilities still apply to the same original total population.

ET	$\mathbb{C}$	$\neg \mathbb{C}$
Death	0.5	0
Life	0	0.5

This only goes to show that the human mind might work best when it can reduce complex reality to simple causality or logic. If we would meet phenomena with that particular reaction as in Model 2 then we would quickly find a new word for the particular combination and include it under the list of dangerous events. A story might be that Papa Mafia flips a coin and Mama Mafia flips a coin, and if they agree then you' re done for. In that case we would call "mob agreement" the risk.

A comment is that we should not suffer from the "confusus simplicitatis". When complex mathematical techniques are illustrated with simple examples then we may pop up with a quicker way to do the simple example. This is very OK as it is. But it should not induce us to think that the mathematical technique would be superfluous.

Appendix G: Return to Kleinbaum et al. (2003) Chapter 10

G.1 Introduction

The body of the text refers to Kleinbaum et al. (2003), the chapter 10 on confounding. It is useful to compare what they do with what this paper does. The main point is that they use the term "confounder" (their book) for the cause (this paper).

G.2 The example problem

The following is the example of chapter 10. We have to re-order the data to get the ETC format. Smoking is the cause of lung cancer while the supposedly toxic chemical is the confounder.

CT[Set, Default, "LC by Smoking or Toxic Chemical"]

		Smokers	¬ Smokers
LC	TCX	26	1
	¬ TCX	12	2
¬ LC	TCX	24	24
	¬ TCX	19	48

(a) Kleinbaum et al. present the story that the researcher assumes that TCX is the cause, but forgets to "control" for the smoking history of the subjects.

Comment: It may indeed happen that the researcher approaches a problem upside down. But is it wise to use that manner to introduce new students to the issue of confounding ?

(b) They say that "smoking is a confounder for the relationship between TCX and cancer".

Comment: In the terminology of this paper, TCX is a confounder since it is a seeming cause but not a true cause. Our terminology is not that "*X* is a confounder since it obscures a relationship between *Y* and *Z*". The latter may be a consequence but is not a defining property.

(c) They say that controlling is "categorizing data to variables".

Comment: The point is clear from the context that one starts with two variables, and then introduces a third. But the given definition is not exact since with three variables, there is controlling in all directions, so all variables become control variables.

(d) Kleinbaum et al. (2003:280): "When we form strata by categorizing the entire dataset according to one or more variables, like smoking history in our example here, we say that we are **controlling** for these variables, which we often refer to as **control variables**. Thus, what looks like a twofold increase in risk when we ignore smoking history, changes in no association when controlling for smoking history. This suggests that the reason why workers exposed to TCX had a twofold increase in risk compared to unexposed workers might be explained simply by noting that there were relatively more

smokers among those exposed to TCX. This is an example of what we call **confounding**, and we say that smoking history is a confounder of the relationship between TCX exposure status and ten-year risk for lung-cancer. In general, confounding may be described as a distortion in a measure of association, like a risk ratio, that may arise because we fail to control for other variables, for example, smoking history, that might be risk factors for the health outcome being studied. If we fail to control the confounder we will obtain an incorrect, or biased, estimate of the measure of effect."

Comment d1: There are three meanings for the term "controlling for a variable": (1) to introduce it, (2) taking a conditional, (3) to set it at some value. The terminology is a bit confusing since there are both mathematics and empirics to consider. For mathematics, the variable suddenly appears on the page, and thus is introduced in the domain of discussion. For empirics, the variable has always been there, since whenever one does an observation then one cannot do so without, say, a smoking history being present (unless one has explicitly taking care of it).

Comment d2: Their statement presumes knowledge about what true risk factors "might be", for which one controls. But it is not explained how one determines a true factor, in the presence of the true confounder.

Comment d3: It would also be better to make a sharper distinction between association (correlation) and the effect measure (in this case relative risk).

Comment d4: The statement by Kleinbaum et al. makes more sense when we consider the problem of two causes. In that case one must control for one cause to find the effect of the other cause. But this seems a less relevant case of confounding. In didactics, one would start with a case where one variable is a seeming cause but a true confounder. As, indeed, Kleinbaum et al. do (but turning the problem upside down). In that case, the true problem is to determine which is the cause and which the confounder.

G.3 Crude and adjusted relative risk

Kleinbaum et al. (2003:281) suggest to compare a "crude" relative risk with an "adjusted" relative risk, where the latter is a weighted average of the relative risks of submatrices, that arise by the introduction of a third variable. This is a "data-based criterion".

- Given the true state, the epidemiologist would have a hard time deciding whether there is confounding, and likely decide, finally, using the Kleinbaum

et al. criterion, that there is no confounding (in their terminology) since the crude relative risk does not really differ from the adjusted one, while the two subset relative risks are not too far apart.

ETCAdjustedRRisk[CT[Data]] // N

{11.7284, 9.67742, 13., 11.2748}

When we reorder the data, then we reproduce the Kleinbaum et al. story-line.

CT[Order, {"Effect", "Exposure", "Smoking"}]

		TCX	¬ TCX
LC	Smokers	26	12
	¬ Smokers	1	2
¬ LC	Smokers	24	19
	¬ Smokers	24	48

- In this case, Kleinbaum et al. state that the crude relative risk of 2 differs too much from the adjusted relative risk of 1.2, so that there is confounding due to the presence of the smoking history. They also observe: “these two stratum-specific risk ratios suggest no association between exposure to TCX and the development of lung cancer".

ETCAdjustedRRisk[%] // N

{2.08286, 1., 1.34333, 1.17827}

Comment 1: Iin the philosophy of the epidemiologists, when the problem is looked at from the angle of the confounder (presuming it to be the cause) then there is confounding and when the problem is looked at from the angle of the cause then there would be no confounding (even though the confounder can be present, in our terms). The relative risk rules would provide a guide to find the right direction. We infer that the kind of confounding that the epidemiologists identify here is the “confusus directionis”. The “confusus magnitudinis” is resolved by presenting the adjusted relative risk as the true summary statistic.

Comment 2: The relative risk rules form only a rule of thumb. The adjusted relative risk is a weighed average of the sub-matrix relative risks, and there is not given any theoretical base for such weighing. The ranges for the relative risks aren’t exact and depend upon expert judgement. Presumably, though, these can be made more exact by testing at levels of significance (though Kleinbaum et al. (2003:293) consider this a validity issue that should not be subjected to significance tests).

Comment 3: When we run the ETCStatistics routine then we find that there is a partial conditional independence. We don’t measure the strength of the deviation from full

conditional independence, but it might well be that a difference in risks between $r = 0.39$ and $r_F = 0.52$ might be seen as not too strong. Thus, it may well be that the hidden criterion used by the epidemiologists is relative freedom or conditional independence. (Just like it is with other epidemiologists.)

Naturally, it depends then as well upon the case at hand what deviation is acceptable. And just to be clear; this present paper didn't formulate an (alternative) rule of thumb; we focussed on just analyzing the case, relying on the causal model.

```
((res = ETCStatistics[CT[Data]] // N) // MatrixForm)
```

Matrix ETCStatistics["Cause, True, Ratio"]

	Cause	¬ Cause	Total
Success	0.148148	0.0246914	0.17284
¬ Success	0.234568	0.592593	0.82716
Sum	0.382716	0.617284	1.

Matrix ETCStatistics["Cause"]

	Cause	¬ Cause	Total
Success	38	3	41
¬ Success	43	72	115
Sum	81	75	156

Matrix ETCStatistics["Confounder"]

	Cause	¬ Cause	Total
Confounder	50	25	75
¬ Confounder	31	50	81
Sum	81	75	156

Matrix ETCStatistics["Seeming"]

	Confounder	¬ Confounder	Total
Success	27	14	41
¬ Success	48	67	115
Sum	75	81	156

$$\left(\begin{array}{l}
N \to 156. \\
\text{NSuccess} \to 41. \\
\text{NCause} \to 81. \\
\text{NConfounder} \to 75. \\
\text{MarginalPr(Success)} \to 0.262821 \\
\text{MarginalPr(Cause)} \to 0.519231 \\
\text{MarginalPr(Confounder)} \to 0.480769 \\
\text{IndependentPr(Truth, Confounding)} \to \text{False} \\
(\text{Success} \perp \neg\,\text{Confounder})(\text{Cause}) \to \text{False} \\
(\text{Success} \perp \neg\,\text{Confounder})(\neg\,\text{Cause}) \to \text{True} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}, \neg\,\text{Confounder}] \to 0.387097 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause}, \neg\,\text{Confounder}] \to 0.04 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}, \text{Confounder}] \to 0.52 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause}, \text{Confounder}] \to 0.04 \\
\text{Risk} \to \begin{pmatrix} 0.52 & 0.04 \\ 0.387097 & 0.04 \end{pmatrix} \\
\text{Interaction} \to \{\text{Add} \to 0.132903, \text{Times} \to 3.32258\} \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Cause}] \to 0.469136 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Cause}] \to 0.04 \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\text{Confounder}] \to 0.666667 \\
\text{ConditionalPr}[\,\text{Cause}\,][\,\neg\,\text{Confounder}] \to 0.382716 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\text{Confounder}] \to 0.36 \\
\text{ConditionalPr}[\,\text{Success}\,][\,\neg\,\text{Confounder}] \to 0.17284 \\
\text{RRisk(True)} \to 9.67742 \\
\text{RRisk(Cause)} \to 11.7284 \\
\text{RelativePr(Confounder)} \to 1.74194 \\
\text{RRisk(Seeming)} \to 2.08286 \\
\text{ETCAdjustedRRisk} \to \{11.7284, 9.67742, 13., 11.2748\} \\
\text{Conditions} \to \{\text{True}, \text{True}, \text{True}, \text{True}, \text{False}\} \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\text{Cause}, \neg\,\text{Confounder}] \to 0.612903 \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\neg\,\text{Cause}, \neg\,\text{Confounder}] \to 0.96 \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\text{Cause}, \text{Confounder}] \to 0.48 \\
\text{ConditionalPr}[\,\neg\,\text{Success}\,][\,\neg\,\text{Cause}, \text{Confounder}] \to 0.96 \\
\text{Safety} \to \begin{pmatrix} 0.48 & 0.96 \\ 0.612903 & 0.96 \end{pmatrix} \\
\text{SimpleCauseQ} \to \begin{pmatrix} \text{False} & \text{False} \\ \text{False} & \text{False} \end{pmatrix} \\
\text{ETCSimpson} \to \{\text{Necessary} \to \text{False}, \text{Sufficient} \to \{\text{True}, \text{True}, \text{False}\}\}
\end{array}\right)$$

G.4 A priori criteria

Next to the "data-based criterion for confounding", Kleinbaum et al. (2003:285) present "a priori criteria": "The first a priori criterion is that a confounder must be a risk factor for the health outcome. (...) The second criterion is that a confounder cannot be an intervening variable between the exposure and the disease. (...) The third criterion is that a confounder must be associated with the exposure in the source population being studied."

Comment 1: Criterion 1 had already been mentioned and found confusing for our terminology that a confounder is a seeming cause. In their terms a confounder would be a cause. They use the word "confounder" consistently for the cause, thus not only in the example, but also for these criteria. Thus in chapter 1 - 9 the student has been using the word "cause" for the cause but suddenly the word "confounder" must be used for it.

Comment 2: Criterion 2 would be contradictory to their criterion 1, since an intervening variable would be on the causal path and be a direct factor.

Comment 3: Their explanation is: "Consider a study to assess whether a particular genetic factor, BRCA1, is a determinant of breast cancer. (...) even if by chance, age turned out to be associated with the gene in the study data, we would not control for age, even though there is data-based confounding, because age does not satisfy all a priori criteria." This means that their third criterion was not formulated exact. They said "association" which we normally read as "correlation" but apparently they mean "causally related" which is criterion 1 again. Nevertheless, this leaves the whole issue of association by correlation unaddressed, which is precisely the big problem in confounding.

Comment 4: They also say (p286): "The main difficulty in assessing the third a priori criterion concerns how to determine the association of the suspected confounder, **C,** with the exposure, **E**, in the (...) source population. This requires some knowledge of the epidemiologic literature about the relationship between **C** and **E** and about the source population being studied." The problem posed here must be understood as an issue of study design. Before collecting the data, one decides what data will be collected. Once the data have been collected, it would seem that the causal model and the ETC analysis helps to determine the direction of causation and the size of the impact. But in the design of the study, it would require attention why one would not measure the causal factors.

G.5 Confounding and interaction

Kleinbaum et al. (2003:289): "Another reason to control variables in an epidemiologic study is to control for interaction. To assess interaction, we need to determine whether the estimate of the effect measure differs at different levels of the control variable." and (p290): "Confounding and interaction are different concepts. Confounding compares the estimated effects before and after control whereas interaction compares estimated effects after control. When assessing confounding and interaction in the same study it is possible to find one with or without the other."

They also say for a particular dataset 2: "It appears that there is a protective effect of exposure on disease in stratum 1, but a harmful effect of exposure in stratum 2. In this situation, the assessment of confounding is questionable and potentially very misleading, since the important finding here is the interaction effect(...)."

Comment: It does not seem that we really can make such a difference between interaction and confounding in this manner. With respect to their (not our) original definition that a confounder mixes up a measure of association, and we may take interaction as such a measure of association, then confounding might mix this up, and then the distinction of "before and after control" cannot be relevant since we have to use the confounder to determine its effect. Rather, it seems that the authors are caught up with their counter-intuitive definition of the term confounder, and that they start to apply notions of cause and effect without reconsidering their terms. PM. Given the different impacts of the cause in the two strata, it would seem that we have two causes here, and not just one cause and one confounder (our terms). Unless our causal model suggests that it is a true confounder indeed so that the impact is merely statistical (which would be proper for a true confounder).

G.6 Additive or multiplicative interaction

The following is without comment but is useful to explain the Interaction in the output of the ETCStatistics routine.

Kleinbaum et al. (2003:291) mention the possibilities of additive or multiplicative interaction, and state the measures of risk difference and relative risk to test for these. In particular they stratify on the confounder (our term), then normalize the risks to the background risk and express the relative risk also as a difference measure.

```
TableForm[lis = {{{a1, a2}, {b1, b2}}, {{c1, c2}, {d1, d2}}},
  TableHeadings → CT["ETC", TableHeadings]]
```

	Cause		¬ Cause
Success	Confounder	a1	b1
	¬ Confounder	a2	b2
¬ Success	Confounder	c1	d1
	¬ Confounder	c2	d2

- These are the P[i, j] probabilities or risks. The background risk is b2 / (b2 + d2). PM. We would need to transpose to get the ETC risk layout (not done).

```
TableForm[pmat = lis[[1]] / (lis[[1]] + lis[[2]]) ,
  TableHeadings → {{C, ¬ C}, {F, ¬ F}}]
```

	F	$\neg F$
C	$\frac{a1}{a1+c1}$	$\frac{a2}{a2+c2}$
$\neg C$	$\frac{b1}{b1+d1}$	$\frac{b2}{b2+d2}$

There is no additive interaction when the Add → term is zero, meaning that the risk differences are equal P[1, 1] - P[1, 0] = P[0, 1] - P[0, 0]. There is no multiplicative interaction when the Times → term is zero. With Ratio[i, j] = P[i, j] / P[0, 0] the multiplicative statistic is Ratio[1, 1] - Ratio[1, 0] Ratio[0, 1]. Alternatively the relative risks are equal P[1, 1] / P[1, 0] = P[1, 1] / P[0, 0]. The output of ETCStatistics uses the latter relative risk difference but below we will use the ratio difference.

```
AdditiveOrMultiplicative222[lis]
```

$$\left\{P[1,\,1] \to \frac{a1}{a1+c1},\ P[1,\,0] \to \frac{a2}{a2+c2},\ P[0,\,1] \to \frac{b1}{b1+d1},\right.$$
$$P[0,\,0] \to \frac{b2}{b2+d2},\ \text{Add} \to \frac{a1}{a1+c1} - \frac{a2}{a2+c2} - \frac{b1}{b1+d1} + \frac{b2}{b2+d2},$$
$$\text{Ratio} \to \left\{\frac{a1\,(b2+d2)}{b2\,(a1+c1)},\ \frac{a2\,(b2+d2)}{b2\,(a2+c2)},\ \frac{b1\,(b2+d2)}{b2\,(b1+d1)}\right\},$$
$$\left.\text{Times} \to \frac{(b2+d2)\left(\frac{a1}{a1+c1} - \frac{a2\,b1\,(b2+d2)}{b2\,(a2+c2)\,(b1+d1)}\right)}{b2}\right\}$$

This reproduces their numerical example.

```
AdditiveOrMultiplicative222[13.3, 4.4, 4.8, 3.4]
```

```
{P[1, 1] → 13.3, P[1, 0] → 4.4, P[0, 1] → 4.8, P[0, 0] → 3.4,
  Add → 7.5, Ratio → {3.91176, 1.29412, 1.41176}, Times → 2.08478}
```

Some definitions may be seen more clearly by using the fully parameterized contingency table.

```
lis = SafetyToETCArray[{c, r, b}, {f, q}, {w, v}];

TableForm[lis, TableHeadings → CT["ETC", TableHeadings]]
```

		Cause	¬ Cause
Success	Confounder	$(c-(1-f)\,q)\,(1-w)$	$(-c+f\,(1-q)+q)\,(1-v)$
	¬ Confounder	$(1-f)\,q\,r$	$b\,(1-f)\,(1-q)$
¬ Success	Confounder	$(c-(1-f)\,q)\,w$	$(-c+f\,(1-q)+q)\,v$
	¬ Confounder	$(1-f)\,q\,(1-r)$	$(1-b)\,(1-f)\,(1-q)$

```
AdditiveOrMultiplicative222[lis] // Simplify
```

$$\left\{\text{P}[1,\,1]\to 1-w,\ \text{P}[1,\,0]\to r,\ \text{P}[0,\,1]\to 1-v,\ \text{P}[0,\,0]\to b,\right.$$
$$\left.\text{Add}\to b-r+v-w,\ \text{Ratio}\to\left\{\frac{1-w}{b},\ \frac{r}{b},\ \frac{1-v}{b}\right\},\ \text{Times}\to\frac{-w\,b+b+r\,(v-1)}{b^2}\right\}$$

We now transpose:

```
TableForm[pmat = lis[[1]] / (lis[[1]] + lis[[2]]) // Simplify // Transpose,
  TableHeadings → {{F, ¬ F}, {C, ¬ C}}]
```

	C	$\neg C$
F	$1-w$	$1-v$
$\neg F$	r	b

Appendix H: A note on the teaching order

Didactics are a personal issue. For example, there are active and passive students, and abstract and concrete thinkers, so that we already have 4 combinations in learning styles. And so on. Yet, there are some issues that might be the same for all when considering these issues on causality and epidemiology.

Colignatus (2007e) starts out from logic and the 2 variable world of $p \Rightarrow q$, exemplified by "If it rains then the streets are wet". For a simple cause this reduces to an equivalence, since there is no third variable that can explain why the streets would be wet when there is no rain. Only by allowing for "causes not mentioned otherwise" or an "error term" then we can have a mere implication. See Colignatus (2007e) how this works.

The question that arises is whether those "other causes" might be confounding or not. In the current set-up they would be real causes, like the city street cleaners who make the streets wet once in a while. The error term occupies a single cell (or a row when they

also clean the streets when it rains). There is no reason to consider this confounding and it exists merely to allow for an implication rather than an equivalence.

Subsequently, there is the introduction of a real third variable. How can this be done in a didactically clear manner ? If we would allow for another cause, then this would require an adjustment of the causal model, and that would complicate things. Thus, the simplest introduction is one of a confounder, that seems like a cause but isn't. Here the example of rain is less fortunate since there are no clear confounders. The example of lung cancer, smoking and toxic TCX is more useful.

```
CT[Set, Default, "LC by Smoking or Toxic Chemical"]
```

		Smokers	¬ Smokers
LC	TCX	26	1
	¬ TCX	12	2
¬ LC	TCX	24	24
	¬ TCX	19	48

Now the question is: how to introduce this case in a didactically clear manner ? Some people might argue that 2 × 2 tables always start with averages, where the data are not "controlled" for the absence of a confounder. This teaching approach would be "realistic" since it confronts the student with what will be the normal starting position in a research where one does not know ahead what the causes and confounders will be. In the same way, one would start Kindergarten with tax bills since this is what the kids will be confronted with later on in life. The alternative approach is student-friendly and starts the discussion with a 2 × 2 table that is conditioned on the absence of the confounder, so that all relations are the causal ones.

```
TableForm[Map[Last, CT[Data], {2}],
  TableHeadings → Drop[CT[TableHeadings], −1]]
```

	Smokers	¬ Smokers
LC	12	2
¬ LC	19	48

```
RRisk[%] // N
```

3.02256

This also clarifies that we no longer have a simple implication $p \Rightarrow q$ or equivalence $p \Leftrightarrow q$. Smoking is a contributing factor and not an exclusive cause for lung cancer. Indeed, if we still had a simple cause, or a real implication with some error term, then the introduction of a third variable might be dubious, since it might be too easy to expose the confounder. (This depends upon finding a good example.)

Above causal model is fully described by two risks: *r* and *b*. Note that this may actually require a fourth variable of "other causes" or an error term, for when both cause and confounder are absent.

```
lis = SafetyToETCArray[{c, r, b}, {f, q}, {w, v}];
```

```
{{ConditionalPr["Success"]["Cause", ! "Confounder"],
    ConditionalPr["Success"][! "Cause", ! "Confounder"]},
  {ConditionalPr[! "Success"]["Cause", ! "Confounder"],
    ConditionalPr[! "Success"][! "Cause", ! "Confounder"]}}
```

$$\begin{pmatrix} \text{ConditionalPr}[\text{Success}][\text{Cause}, \neg\,\text{Confounder}] & \text{ConditionalPr}[\text{Success}][\neg\,\text{Cause}, \neg\,\text{Confou} \\ \text{ConditionalPr}[\neg\,\text{Success}][\text{Cause}, \neg\,\text{Confounder}] & \text{ConditionalPr}[\neg\,\text{Success}][\neg\,\text{Cause}, \neg\,\text{Confo} \end{pmatrix}$$

```
% /. ETCStatistics[lis, Print → False]
```

$$\begin{pmatrix} r & b \\ 1-r & 1-b \end{pmatrix}$$

The subsequent step is to reduce abstraction and introduce more realism by including the data set with the confounder present.

This also introduces the averages for *R,* ***B,*** R_F and B_F.

Steps are: (i) $r = 1$, $b = 0$, $w = 1 - r$, $v = 1 - b$, (ii) conditional independence, (iii) marginal independence, (iv) none of these, with more or less safety, due to random effects, (v) the difficult issue of *q* and *f* as causes or correlations, (vi) errors in measurement such that the causal relation becomes more blurred, (vii) somewhere along the line: plug in the Simpson paradox when it might start to bite. As soon as one allows for more scope for random effects the approach would be less causal and more statistical, and then an answer would be to allow more observations in different contingency tables over time to recover both the causality and the distributions.

Once these issues are understood, then one may pose the question as to what happens in the reverse, when we have only the averages to start with, and then run a randomized controlled trial to find the inner matrix.

After this is understood, one might discuss different research formats, such as disease-test, treatment-control, case-control, survival analysis, etcetera. Probably the best format to start with is not lung cancer, smoking and toxic issue but the disease-test matrix, since this links up with proof theory in logic and hypothesis testing in statistics.

Appendix I: Routines

This discussion uses The Economics Pack, Cool (2001).

? ETCArrayQ

ETCArrayQ[x] returns True if x is a {2, 2, 2} array, and otherwise False

? SafetyToETCArray

SafetyToETCArray[{c, r, b}, {f, q}, {x, y}, n:1] is an
application of ToETCArray so that {x, y} is interpreted as {w, v}
SafetyToETCArray[] contains in Hold how the output of
ETCStatistics could be used to create the same input array

? RiskToETCArray

RiskToETCArray[{c, r, b}, {f, q}, {x, y}, n:1] is an
application of ToETCArray so that {x, y} is interpreted as {R, B}
RiskToETCArray[] contains in Hold how the output of
ETCStatistics could be used to create the same input array

RiskToETCArray[{c, r, b}, {f, q}, {x, y}, n:1] is an
application of ToETCArray so that {x, y} is interpreted as {R, B}
RiskToETCArray[] contains in hold how the output of
ETCStatistics could be used to create the same input matrix

? ToETCArray

ToETCArray[{c, r, b}, {f, q}, {x, y}, n:1] gives a 2 x 2 x 2 table with the
order Effect, Truth, Confounding. Option ETCParms determines how x
and y are interpreted. Method −> Safety is default, the alternative is Risk:
ToETCArray[Safety, {c, r, b}, {f, q}, {w, v}, n:1]
ToETCArray[Risk, {c, r, b}, {f, q}, {R, B}, n:1]
The meaning of the parameters is
c = Pr[C] = marginal of the cause = p f + q (1 − f);
f = Pr[F] = marginal of the confounder
p = Pr[C | F] = chance that the cause occurs given F (solved from c)
q = Pr[C | !F] = chance that the cause occurs given F
r = Pr[S | C, !F] = risk
b = Pr[S | !C, !F] = background risk
R = Pr[S | C] = average risk
B = Pr[S | !C] = average background risk
w = Pr[Not[S] | C, F] = miraculous = wunderbar
v = Pr[Not[S] | Not[C], F] = background safety
e = 1 − r = exceptional safety (the cause's failure rate)
a = 1 − b = background safety (all absent)
n = total number of cases
Option "Round" controls rounding, default it does.
Substitute p = q = c iff c and f are marginally independent.
ToETCArray[Confounder, {c, r, b},
{p, q}, {R, B}, n:1] uses f = (c − q)/(p − q) for p <> q.
Enter r > b otherwise reverse the definition of what is the cause, absence rather than presence.
But R > B is not tested though this essentially would also better relabelled if it is not so.
One would use p > q as well.

? ETCAverageRisksFromSafety

ETCAverageRisksFromSafety[{c, r, b}, {f, q}, {w, v}] gives {R, B}.
w = P[Not[S] | C, F], w from wunderbar = miraculous
v = P[Not[S] | Not[C], F] safety (from Dutch "veiligheid")
See ETCStatistics and ToETCArray for the other parameters

? ETCStatistics

ETCStatistics[mat] for a 2 x 2 x 2 array with the ETC layout of Effect, Truth,
Confounding, generates summary statistics. See ETCPrTable and ETCTable.
ETCStatistics[opts] uses default CT["ETC", Data]
Subtables and probabilities in output are
ETCStatistics["Cause"] for the average cause for Success, with
R = Pr[Success | Cause]
B = Pr[Success | Not[Cause]].
ETCStatistics["Confounder"] for the confounding probabilities
p = Pr[Cause | Confounder]
q = Pr[Cause | Not[Confounder]].
ETCStatistics["Seeming"] for the erroneous
view where the success is related to the confounder, with
Rf = Pr[Success | Confounder]
Bf = Pr[Success | Not[Confounder]]
Option Print −> (default True) controls printing of subtables, Simplify −> (default True)
simplifies tables in printing, N −> (default True) applies to the rate table in printing

? ETCRiskTable

ETCRiskTable[res] uses res = ETCStatistics[...] to make
a table of the various risks and their values, difference and ratio's
ETCRiskTable[res] uses res = Results[ETCStatistics] that are
default present after the statistics routine has been called

? ETCPrTable

ETCPrTable[] lists the table of key risks
and probabilities of the Effect, Truth, Confounding table

? ETCTable

ETCTable[22, lab] takes CT[lab, Data] that
must be 2 by 2, adds border sums, and prints with TableForm
The following are border matrices:
ETCTable["ET", c, {R, B}] for marginal
probability of the cause c, risk R and background risk B
ETCTable["EC", f, {Rf, Bf}] for the marginal probability of the
confounder f, seeming risk Rf and seeming background risk Bf
ETCTable["TC", f, {p, q}] for the marginal probability of the confounder
f, p = Pr[cause | confounder] and q = Pr[Cause | !confounder]
ETCTable["−S|TC", {c, r, b}, {f, q}, {w, v}] shows the table for truth and confounding,
given that there is no success, with w and v the strong and weak safety

? ETCSquare

ETCSquare[] shows ETCSquare[CT[Data]]
ETCSquare[string] shows ETCSquare[CT[string, Data]]
ETCSquare[lis] shows the 2 x 2 x 2 square. For the ETC model, the effect
is in the rows, the truth in the columns and the confounder in the triangle
ETCSquare[Label] just shows the labels
Options are:
Show –> (default True) show the labels
N –> {positions} for the positions of x and the labels
Position –> {a, b} for the position when there are no labels. A position is just a
pair of numbers between 0 and 1, because the other positions are symmetric
TableHeadings –> (default Automatic) the list of labels, and if Automatic, it uses:
First –> {3 strings} for labels of an event
Last –> {3 strings...} for labels of the absence of the event

? ETCSimpson

ETCSimpson[{c, r, b}, {f, q}, {w, v}] gives the necessary and sufficient conditions for the
Simpson paradox to arise, by regarding the F and !F groups for which the risk difference >
0 while the total has risk difference < 0. NB. The output terms are in quotes. NB.
The sufficient conditions are all True if the paradox occurs (only for the stated case)

? ETCAdjustedRRisk

ETCAdjustedRRisk[Safety, {c, r, b}, {f, q}, {w, v}] or
ETCAdjustedRRisk[Risk, {c, r, b}, {f, q}, {R, B}] or
ETCAdjustedRRisk[lis222] give the relative risk {R/B, r/b, RR, (1 − f} r/b + f RR},
where R/B is the crude measure, r/b is the relative risk of the not−F population,
RR of the F population, and their weighted average is the final adjusted measure

Literature

Colignatus is the name of Thomas Cool in science.

Cool, Th. (1999, 2001), “The Economics Pack, Applications for *Mathematica*”, http://www.dataweb.nl/~cool, ISBN 90-804774-1-9, JEL-99-0820

Colignatus, Th. (2007d), “Correlation and regression in contingency tables. A measure of association or correlation in nominal data (contingency tables), using determinants", http://mpra.ub.uni-muenchen.de/3394/ Retrieved from source

Colignatus, Th. (2007e), "Elementary statistics and causality", work in progress, http://www.dataweb.nl/~cool/Papers/ESAC/Index.html

Colignatus, Th. (2007f), "The 2 × 2 × 2 case in causality, of an effect, a cause and a confounder", http://mpra.ub.uni-muenchen.de/3351/, This paper. Retrieved from source

Colignatus, Th. (2007g), "A comparison of nominal regression and logistic regression for contingency tables, including the 2 × 2 × 2 case in causality", to appear

Fisher, R.A. (1958a), "Lung Cancer and Cigarettes? Letter to the editor", *Nature*, vol. 182, p. 108, 12 July 1958 [*Collected Papers* **275**], see Lee (2007), http://www.york.ac.uk/depts/maths/histstat/fisher275.pdf, Retrieved from source

Fisher, R.A. (1958b), "Cancer and Smoking? Letter to the editor", *Nature*, vol. 182, p. 596, 30 August 1958 [*Collected Papers* **276**], see Lee (2007), http://www.york.ac.uk/depts/maths/histstat/fisher276.pdf, Retrieved from source

Kleinbaum, D.G., K.M. Sullivan and N.D. Barker (2003), "ActivEpi Companion texbook", Springer

Lee, P.M. (2007), "Life and Work of Statisticians", http://www.york.ac.uk/depts/maths/histstat/lifework.htm, Revised 24 April 2007

Pearl, J. (1998), "Why there is no statistical test for confounding, why many think there is, and why they are almost right", UCLA Cognitive Systems Laboratory, Technical Report (R-256), January 1998

Pearl, J. (2000), "Causality. Models, reasoning and inference", Cambridge

Saari, D.G. (2001), "Decisions and elections", Cambridge

Schield, M. (1999, 2003), "Simpson's paradox and Cornfield's conditions", Augsburg College ASA-JSM, http://web.augsburg.edu/~schield/MiloPapers/99ASA.pdf, 07/23/03 Updated, Retrieved from source